MESP1's EPIGENETIC AND TRANSCRIPTIONAL ROLE DURING MESENDODERM FORMATION

_____

A Dissertation Presented to

the Faculty of The Department of Biology and Biochemistry

University of Houston

_____

In Partial Fulfillment

of the Requirements for The Degree

Doctor of Philosophy

_____

By

Jong Hwan Kim

December 2016

MESP1's EPIGENETIC AND TRANSCRIPTIONAL ROLE DURING

MESENDODERM FORMATION

_____

Jong Hwan Kim. Student

APPROVED:

_____

Robert J. Schwartz, Ph.D.

Committee Chair

_____

Amy Sater, Ph.D.

Biology and Biochemistry Chair

_____

Preethi Gunaratne, Ph.D.

_____

Bradley K McConnell, Ph.D.

College of Pharmacy, University of Houston

_____

Dan Wells, Ph.D.

Dean, College of Natural Sciences and Mathematics

Department of Biology and Biochemistry

ACKNOWLEDGEMENTS

I would like to give special thanks to my primary investigator, Dr. Robert J. Schwartz for supporting me throughout my studies and giving me the opportunity to develop my investigative skill within his lab. He is a very kind and whole-hearted person who has made a significant impact in how I will conduct myself in my professional career. He made my time here at the University of Houston very memorable and enjoyable. I cannot thank him enough. I would also like to personally thank Dr. Preethi Gunaratne who has been a mentor since the fall of 2008 when I finished my undergraduate degree. She allowed me to volunteer in her lab and the opportunity she provided me was essential for opening doors to the path that I'm currently embarking on. Without her trust and faith in me, I do not know where I would be right now. I would also like to thank Dr. Amy Sater for her support and interest in my investigation and for providing guidance during committee meetings. Her passion for developmental biology encouraged me to investigate the developmental impact on my study diligently. I would also like to thank Dr. Bradley McConnell for accepting to be part of my committee midway through my studies. I would also like to thank the staff, and fellow investigators present and not present in Dr. Schwartz lab with their patience for all my inquiries, Dr. Yu Liu, Dr. Li Chen, Dr. David Stewart, Dr. Vladimir Potaman, and Dr. Benjamin Soibam. Your insight in your respective fields was essential and critical for the development of my investigation. I would also like to thank Dr. Ashley Benham who recruited me to work for Dr. Schwartz and her trust in my qualifications. I would like to thank Dr. Kuo-Chan Weng for aiding me and working alongside me in our

study of MESP1. I would like to thank all my friends and colleagues in and out of the lab that made my time at the University of Houston extremely enjoyable, and special thanks to Nicolas Valenzuela for some very good times.

MESP1's EPIGENETIC AND TRANSCRITIONAL ROLE DURING MESENDODERM FORMATION

_____

An Abstract of a Dissertation

Presented to

the Faculty of the Department of Biology and Biochemistry

University of Houston

_____

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

_____

By

Jong Hwan Kim

December 2016

ABSTRACT

MESP1 is a basic helix loop helix transcription factor that is essential for the survival and development of mouse embryogenesis. It is the earliest marker for identifying the nascent mesoderm that is fated to becoming the myocardium, head mesenchyme, and somites. In spite of the advancements in our knowledge of MESP1, it is, however, still very unclear how it directs the activation of the cardiac program. It is important to understand MESP1's epigenetic, and transcriptional function during development to increase efficiency in somatic and stem cells cardiac reprogramming. ChIP-seq analysis of endogenous targets of endogenously expressed MESP1 affected neuro-ectoderm specific GO terms over mesendoderm specific terms predominantly. To analyze MESP1's effect on target genes, a comparison to RNA-seq data of FACsorted YFP cells from ESC and day 5, 6, 7, and 8 was done. YFP protein marks endogenously activated MESP1 expressing cells during EB differentiation which resulted in identifying MESP1's effect on target genes. Analysis of the transcriptome of YFP positive cells provided evidence that MESP1 is indirectly activating the core cardiac program and directly repressing the non-mesoderm program such as neuro-ectoderm. MESP1 directly represses the neuro-ectoderm developmental program by targeting, Sox2, Neurod1, Neurog1, Neurog2, and Neurog3 by reducing the expression levels within the first 12 hours of Dox induction. Meta-data analysis of MESP1-binding regions with H3K27acetylation and tri-methylation show strong overlap with H3K27acetyaltion favoring Mesendoderm specific genes and H3K27me3 favoring neuro-ectoderm target genes. H3K27acetylation suggest a potential mechanism

in indirectly activating the core cardiac program. ChIP-qPCR of *Neurog3* adjacent enhancers show that MESP1 guides the deposit of H3K27me3 through the PRC2 complex in an ebox variant-dependent manner. Analysis of MESP1 binding sites for potential bias in the ebox variant show that CACCTG variant is favored in regards to both repressive function and H3K27me3 marking,.

CONTENTS

1. Chapter 1 Literature Review

The discovery of Mesp1 in 1996 by Dr. Yumiko D. Saga from the Banya Tsukauba Research institute in Ibaraki Japan has open the door for decades of research in understanding the developmental process of cardio-myogenesis (Saga, Hata et al. 1996). More than over 800 peer-reviewed articles have been published on MESP1 biogenesis, regulation, and obligatory role in the formation of cardiac lineages. Thus, Dr Saga's seminal discovery was an important achievement in contemporary developmental biology. Recent studies also revealed that populations of Mesp1-marked ES cells may hold the key to understanding how embryonic stem cells become cardiac progenitor lineages. There is a great urgency today to replace damaged cardiomyocytes with new myocytes from reprogrammed adult mesenchymal stem cells and or somatic cells. Thus, understanding MESP1's functional biological role is essential for generating new insights in regenerative medicine.

Cardio-myogenesis and the developmental appearance of the heart begins during the earliest stages of embryogenesis being the first mesodermal organ to appear after gastrulation. MESP1 and its importance in embryogenesis, requires an understanding of gastrulation and the formation of the mesodermal germ layer.

Gastrulation is the process of producing the three germ layers from the single-layer blastula. Each germ layer (ectoderm, mesoderm, and endoderm) further develops to give rise to distinct tissues of the organism. The initiation of gastrulation is marked by the formation of the primitive streak, located on the

surface of the blastula towards the posterior regions of the embryo (Read, Bedford et al. 1970, Kennison 2002, Wild and Fleming 2002). Mesoderm formation begins at the junction of the epiblast and the extraembryonic tissue where Dr. Saga discovered the appearance of Mesp1. The primitive streak extends down the midline, establishing both the bilateral symmetry and posterior-anterior axis, and the ingression of the mesoderm and endoderm progenitors begins and the formation of the three germ layers occurs (Kinder, Tsang et al. 2001).

When the three germ layers are established, they will begin to elaborate into distinct parts of the organism. The ectoderm germ layer will further differentiate into the surface ectoderm, neural crest, and neural tube (Tuchmann-Duplessis, David et al. 1971, Stern 2005). The germ layer endoderm contributes to all the inner epithelial lining of the digestive tube, and the lining of all glands of the digestive tubes, and stomach, colon, liver, pancreas, and the urinary bladder. The endoderm also contributes to the epithelial lining of the trachea, lung, pharynx, thyroid, parathyroid, and intestines (Wells and Melton 1999, Kimelman and Griffin 2000, Rodaway and Patient 2001, Hogan and Zaret 2002, Technau and Scholz 2002, Zorn and Wells 2009, Goessling and Stainier 2016). The mesoderm layers that lies between the ectoderm and the endoderm, generates into intermediate mesoderm, paraxial mesoderm, lateral-plate mesoderm, and chorda-mesoderm; in which contributes to muscle, bone, cartilage, connective tissue, adipose tissue, circulatory system, lymphatic system, dermis, genitourinary system, serous membranes, and the notochord (Tuchmann-Duplessis, David et al. 1971, Kimelman and Griffin 2000, Technau and Scholz 2002, Papaioannou 2004,

Devine, Wythe et al. 2014, Kyba 2016). The focus of this study will be on the Mesp1 role in defining mesoderm formation, due to its impact on the formation of cardiomyocytes.

MESP1 is 243 amino acids in length and contains a basic helix loop helix domain characterized by analyzing the isolated cDNA and genomic sequence. MESP1 mRNA transcripts were observed in the allantois region between the extraembryonic, and the epiblast of a 7.5 days post coitum (dpc) mouse embryo and validation was visualized using whole-mount *in-situ* hybridization (Figure 1.2) (Saga, Hata et al. 1996).

The spatial patterning of MESP1 was also compared to alkaline phosphatase (ALK), marking the primordial germ cells (PGC). The first sign of MESP1 expression occurred at around 6.5 to 6,75 dpc at the junction of the epiblast and extraembryonic ectoderm. Saga and colleagues noted that Mesp1 expression was quickly down-regulated by 7.5 dpc in the mesodermal cells that begin migrating towards the anterior region, but MESP1 transcripts remain at the base of the allantois until 8.5 dpc. MESP1 was not located past the embryonic node during lateral migration. Transverse section of 6.5 dpc embryos showed that MESP1 RNA was expressed in embryonic mesoderm but not in the ectoderm or endoderm.

In 1998, Dr. Saga demonstrated that MESP1 was necessary for proper morphogenesis of the heart by generated *MESP1*-deficient embryos. She found that MESP1-null mouse embryos exhibited growth retardation by 7.5 dpc and

death by 10.5 dpc due to abnormal head, heart, and somite formation (figure1.3A-D) (Saga 1998, Saga, Miyagawa-Tomita et al. 1999). MESP1-null mice exhibited underdeveloped small head folds, neural fold closure, and embryonic turning was delayed (Figure1.4C, D). The null embryos had two symmetrical heart tubes, each with the ability to beat suggesting that partial cardio-myogenesis still occurred. Transverse sectioning showed two completely separated heart tubes on both sides of the midline (Saga 1998).

Lineage tracing of *MESP1*$^{lacZ/+}$ and *MESP1*$^{lacZ/-}$ embryos showed strikingly different patterning during development. In *MESP1*$^{lacZ/+}$ embryos, β-gal stained cells showed a lateral migration from the primitive streak, reaching the anterior regions. β-gal staining of a cross-section of the heart also showed evidence that MESP1 cells contributed to and confined to the myocardium, (Figure1.5D, E) However, the *MESP1*$^{lacZ/-}$ did not undergo lateral migration from the primitive streak but remained. (Saga, Miyagawa-Tomita et al. 1999).

Dr. Saga concluded that the major defect from *MESP1*-null mice was reduced migration of the mesodermal cells. She also concluded that MESP1 is the earliest molecular marker for identifying heart precursor cells. MESP1 appears to function in other mesodermal lineages, the *MESP1*-null embryos showed the delayed generation of the paraxial mesoderm, but was rescued by MESP2 (Saga 1998).

Dr. Saga and her colleagues showed that MESP1 has a critical role in embryogenesis and morphogenesis of the heart field which was supported by

extensive lineage tracing which determined the fate of MESP1 marked cells (Saga, Hata et al. 1996, Saga, Hata et al. 1997, Saga 1998, Saga, Miyagawa-Tomita et al. 1999, Kitajima, Takagi et al. 2000, Saga, Kitajima et al. 2000, Takahashi, Hiraoka et al. 2005, Kitajima, Miyagawa-Tomita et al. 2006, Morimoto, Kiso et al. 2006), the progression naturally moved towards understanding MESP1's direct role through gain of function investigations.

In 2008, Dr. Theresa L. Murphy and her colleagues discovered that MESP1 could induce the Epithelial-Mesenchymal transition (EMT) and restrict the cardiovascular fate in differentiating embryonic stem cells. This was the first look into what MESP1 could potentially regulate. Dr. Lindsley established that in transiently expressing MESP1 cells, levels of PDGFRa and Flk1 markers of mesoderm germ layer were increased independently of any wnt-signal induction. They also tested T, Cdx2, Eomes, Evx, Hand1, Mixl1, and found that MESP1 was the only factor that could induce the mesoderm program without wnt-signalling (Lindsley, Gill et al. 2007).

MESP1 strongly regulates epithelial-mesenchymal transition (EMT) measured by the increase of E-cadherin and the decrease of N-cadherin seen by immunostaining and by the expression of Snai1. Lindsley et al. (Lindsley, Gill et al. 2007) demonstrated that MESP1 induced the expression of N-cadherin, the key marker for mesenchymal cells and reduced the expression of E-cadherin, the molecular marker for epithelial cells by transiently expressing it in differentiation embryonic stem cells (ESC).

Global microarray analysis on transcripts after 96 hours of induction revealed evidence of lineage restriction capability of MESP1. They found that MESP1 showed potential to inhibit gene activity associated with the neuroectoderm, paraxial mesoderm, and the hematopoietic program. The neuroectoderm and paraxial mesoderm markers used in this study were *Sox2, Pax6,* and *Pax7*. These markers were only expressed when the cells were treated with DKK1 but inhibited by MESP1 with or without DKK1 (Figure 1.9A, B). MESP1 also repressed genes associated with the hematopoietic lineage, but only under the conditions of induction of the Wnt signaling pathway, such as *Tal1, Gata1,* and *hbb-b1*. Also MESP1 can induce cardiac specific genes *Myh6, Myh7, Myl2, Myl7, Tnnt2,* and *Nppa*, but only with DKK1 (Lindsley, Gill et al. 2007).

After eight days of post induction, MESP1 sustained the expression of key cardiac and smooth muscle transcription factors, *Gata4, Gata6, Hand1, Hand2, Mef2c, Smyd1 (Bop), Tbx5, Tbx20 Myocd, Foxc1, Foxc2, Tbx1, Zfhx1a, and Nkx2.5*. In the presence of DKK1, *Nkx2.5* appeared to be preferentially expressed, which corresponds with their finding for contractility and cTNT expression with DKK1 (Lindsley, Gill et al. 2007).

Therefore, MESP1 can increase the conversion of mouse embryonic stem cells into immature smooth muscle cells and functional cardiomyocytes that require WNT inhibition by DKK1. Lindsley et al. (Lindsley, Gill et al. 2007) were able to show that two markers, alpha-smooth muscle actin (aSMC) and cardiac troponin (cTNT) increased during mouse ES cell programming into cardiomyocytes. Within six

days, MESP1 expression increased the presence of aSMC from 4% to 45% in the cell population that didn't display contractile activity. However, with the inhibition of WNT by DKK1, expression of MESP1 was was able to induce cTNT and the cells showed contractile capability (Lindsley, Gill et al. 2007)

Since MESP1 expression *in-vivo* is transient, many of the contractile protein genes such as the myosin heavy chain (MHC), which appear several days later, are likely the result of indirect activation by downstream factors. The short-term effects of MESP1 in mouse-embryonic stem cells at 6 h, 12 h, and 24 h post-Dox induction showed that a third of the activated genes were primarily involved in gene regulation and signal transduction. Within the first six hours, key factors that regulate EMT and cardiovascular lineages were induced, *Snai, and Lhx1.* Thus, according to Lindsley et al. 2007, MESP1 is directly inducing the EMT program, but indirectly initiating the core cardiac program. They also found that Mesp1 inhibited non-mesodermal lineages, as previously shown by the Saga studies (Saga, Hata et al. 1996, Saga, Miyagawa-Tomita et al. 1999, Saga, Kitajima et al. 2000, Lindsley, Gill et al. 2007).

Dr. Antoino Bondue in 2008 within the same year confirmed Dr. Lindsley's findings in MESP1's ability to activate the core cardiac program, *Hand2, Myocardin, Nkx2.5, Gata4, Mef2c, Tbx20, or FoxH1*. He suggested that due to quick response to MESP1, they are direct targets which were not observed by Lindsley et al. (Lindsley, Gill et al. 2007). Bondue also observed that MESP1 inhibited the appearance of alternative cell fates by repressing early primitive streak

specification factors *Brachyury* and *FGF8*. Bondue also showed that MESP1 may repress endoderm gene activity with the inhibition of *Foxa2, Sox17, Gsc, Nodal,* and *Cer1.* While Bondue's temporal study was similar to Lindsley et al. (Lindsley, Gill et al. 2007). There are many discrepancies found between these two studies; thus begging the question what are the MESP1 direct gene targets? Our answer was recently shown in Soibam et al. (Soibam, Benham et al. 2015) and Chapter 2.

There is a clear difference between the time MESP1 is expressed and when the core cardiac program peaks. In EB differentiation MESP1 begins expressing as early as day 3 and is quickly down-regulated by day five, while the core cardiac markers, *Nkx2.5, Hand2, Mef2c, and Troponin*, begin expressing at day 6 and peaks at day 7-8 (Bondue, Lapouge et al. 2008) and a clear separation In Dr. Lu's investigation of MESP1 marked EYFP+ cells, during early mesendoderm activating factors , Eomes, Mixl1, and Mesp1 from the core cardiac program Nkx2-5, Mef2c, and Tbx5 (Liu, Chen et al. 2016).

This gap in time would suggest that MESP1 is more likely indirectly activating the cardiac program, and should be further investigated. The effects of MESP1 is sustained within the population after removal of stimuli suggesting a priming role (Lindsley, Gill et al. 2007). The sensitivity of mES cells to MESP1 is evident which makes it difficult to determine MESP1's direct targets correctly. Also, the artificial expression has the potential of producing false positives. To correctly determine MESP1's valid targets, a global analysis by chromatin immunoprecipitation of MESP1 binding sites of endogenous expression should be

done. This will give us an accurate assessment of MESP1s role during gastrulation and the possible programs that MESP1 activated and repressed (See Chapter 2).

With that being said, the ability for MESP1 to activate the cardiac program autonomously is evident, but the timing between the two events would suggest that MESP1 is indirectly activating the cardiac program. If MESP1 is not directly activating the cardiac program, its other potential role is altering the chromatin to allow other factors to bind. MESP1 potentially could be aiding in chromatin remodeling. Understanding MESP1 chromatin remodeling effects will allow for a finer tuned approach to cellular reprogramming. However what remains to be shown is the mechanism in which MESP1 relays that program after its degradation. Two among many possibilities are 1) Is MESP1 priming the binding sites through chromatin remodeling similar to a pioneer factor, or 2) Recruiting co-factors that will relay the activation signal until target gene is required during differentiation. In my thesis I focused on whether MESP1 is acting as a pioneer factor and altering the epigenetic state through chromatin remodeling.

Epigenetics through chromatin remodeling is the process in which nucleosome DNA is made accessible or inaccessible through changes in structure, composition, and nucleosome positioning. Two modes of remodeling currently have been shown to occur; Covalent histone-modification, and ATP-depending remodeling (Teif and Rippe 2009).

Histone post-translational modification is the addition or removal of various elements from histone enzymes like acetylation, methylation, phosphorylation, and

ubiquitination. The histone family of protein has five distinct variants H1/H5, H2A, H2B, H3, and H4. Each variant except H1/H5 dimerizes and form a complex known as a nucleosome. Each histone variant contains an N-terminal 'tail' that is not part of the tertiary structure and varies in length depending on histone (Strahl and Allis 2000).

The tail of each histone has been the focus in determining what modifications are present and how it affects chromatin remodeling. The complexity of this mechanism is exponential and not yet completely understood. Each histone tail contains multiple residues including lysine, arginine, serine, and threonine, which can be modified independently, synergistically, and differentially for function. Studies into histone post-translational modification addition of methyl, acetyl, and phosphoryl groups onto residues mentioned above can mark for either activation and or repression, euchromatin, heterochromatin, active promoters, active enhancers, and or active transcription (Strahl and Allis 2000).

Since the discovery of histone modification, histone 3 has been the most studied within the complex. Histone3s N-terminal tail that is not part of the globular domain is 38 amino acids long, with 16 known sites for modification that are essential for remodeling. The histone H3 tail contains seven lysine residues that can be acetylated or methylated, four arginine groups that can be methylated or citrullinated, two serine, and two threonine, that can be phosphorylated (Strahl and Allis 2000, Bannister and Kouzarides 2011, Venkatesh and Workman 2015). The combinatorial complexity is significant to understand how genes are being

regulated, but actual genetic loss and gain of function approaches have been successful in identifying specific roles for each modification.

Histone H3 Lysine residue 27 (H3K27), when modified by methylation or acetylation, generates epigenetic marks that correlates with active or repressed gene activity. Histone Lysine 4 (H3K4), when modified by multi methylations, marks active target promoter and enhancer regions. Histone H3 Lysine 36 (H3K36), when methylated, marks the progression of polymerase II transcription. Histone H3 Lysine 9 (H3K9) methylated marks indicate that the target region is poised to be converted to be inactive and or heterochromatin in nature. While other histones have been studied the core features of H3 modifications have been utilized to determine the state of the chromatin in the global epigenetic analysis; thus is important in the analysis of MESP1 expressing cells (Strahl and Allis 2000, Bannister and Kouzarides 2011, Venkatesh and Workman 2015, Fiszbein, Giono et al. 2016). For this thesis, I will focus on MESP1's ability to modulate the acetylation and methylation of Histone 3 lysine 27 (H3K27). This lysine, when modified, have epigenetic implications for transient, quick, transcriptional regulation of target gene. To measure more permanent effects, different modification will need to be investigated (Strahl and Allis 2000).

Each marker is catalyzed by specific core enzymatic complexes and can be potentially guided by transcription factors like MESP1 (Smale 2010). For lysine to be acetylated, a complex called histone acetyltransferase (HAT) must be recruited and bound for acetyl-CoA to be transferred on target lysine. Also, the reverse can

occur in which a histone deacetylase (HDAC) can be recruited and remove acetylation from target lysine. The same occurs for methylation; to add methyl groups to lysine, complex proteins called Histone methyltransferases (HMT) are recruited, for demethylation Histone Demethyltransferases (KDM) are recruited (Strahl and Allis 2000, Bannister and Kouzarides 2011). However, for lysine 27 there are specific HATs, HDACs, HMTs, and KDMs that are involved in the addition and removal of post-translational modification for the specific function of MESP1 (Bannister and Kouzarides 2011).

In 2003, the end of the human genome project, began the revolution of next generation sequencing (NGS) and the technologies surrounding it. It dramatically changed in the way we approach and theorize the way biology systems function (Goodwin, McPherson et al. 2016). Currently, sequencing technologies and upstream techniques like ChIP, ChiRP, DNAse, and ATAC allow for the study of DNA in respects to chromosomal state (histone modification, DNA methylation, etc.), and total RNA in respects to global expression level of micro RNA, Long non-coding RNA, messenger RNA, etc.

Since then, the platforms also evolved to be more accessible and affordable exponentially increase the generation of NGS data. This explosion of information in the past decade fueled the development of new emerging fields of bioinformatics and predictive biology.

Predictive biology or systems biology is the discipline of explaining biologic phenomenon at a cellular holistic approach, through net-interactions of all

components within it. The dawn of NGS that digitized DNA and RNA sequences provided the massive datasets required to successfully predict de novo biological outcomes. This allows for the possibility to integrate biological and analytical systems to generate datasets to predict dynamic cellular processes. However currently bottlenecks in predictive biology is the bridge between the two systems. The issues are focused on analyzing and comparing massive datasets and integration of different databases (Liu 2005).

The processing of NGS datasets to be useful for predictive biologist, requires an understanding of not just the biological dynamic process but the tools in making the dataset useful. In the field of bioinformatics, processing of datasets can either be systematic or subjective. Systematic meaning raw data to useful datasets, for example raw ChIP data processing into peak location, adjacent gene identification using standard software to compile, align, and internal normalization. Subjective analysis is more difficult due to its dependence on understanding the biological implication of each data set and how they should be compared. This is an interdisciplinary hurdle in the development of bioinformatics and predictive biology.

The systematic processing of NGS data of both RNA and DNA require different repertoire of software. However, early portions of the established pipeline are similar since the data sets would need to be aligned to a reference genome. In this study, I used TopHat to align the ChIP-seq and RNA-seq data to the mouse reference genome from the National Center for Biotechnology Information (NCBI).

For the ChIP-seq analysis of Histone modification and MESP1 binding sites I use HOMER software (Heinz, Benner et al. 2010) to annotate each peak with the nearest transcription start site, generate bedgraphs to be visualized on the University of Santa Cruz Genome Browser, and determine the density of each peak. For RNA-seq analysis I use Cufflinks to quantify alignments by Reads Per Kilobase of transcript per Million mapped reads (RPKM). Up to this point, it would be possible for an individual with little to no biological background to process and organize NGS datasets. After this point in the pipeline is highly subjective and requires the interdisciplinary coordination that I mentioned earlier and the hurdle that is impeding predictive biology.

The subjective aspect of bioinformatics requires an understanding of both the biological experimental process of unknown vs controls and the tools available for data processing, or an efficient communication between the experimental biologist and the bioinformatictian. This limit is extremely difficult to overcome and solutions to increase productivity would require a generation of investigators trained enough in both disciplines. Also it would be difficult to expect computational sophistication from a generation of biologists and vice-versa (Liu 2005). The current solution of increasing productivity is to simplify the systematic process by developing web-based user-friendly interfaces of well-established programs like tophat (Kim, Pertea et al. 2013), bowtie (Langmead, Trapnell et al. 2009), and HOMER (Heinz, Benner et al. 2010) to allow experimental biologist to generate useful dataset, circumventing the need for highly skilled bioinformatics or computer

scientist. This trend will however meet its limit due to the limitation of current available programs.

My investigation of MESP1 and its epigenetic and transcriptional role will highlight both the systematic and subjective process of bioinformatics and the potential resourcefulness of predictive biology or systematic biology. With Chromatin immunoprecipitation of MESP1, Histone Modification, and RNA-seq data, the effects of MESP1 is predicted by cross-comparison of different data-set which then are tested experimentally.

We hypothesize that during expression of MESP1, its primary role may include guiding specific enzymatic complexes to designated sites to add and or remove specific histone modifications for activation and repression. Understanding this role may lead us to understand whether MESP1 is indirectly activating the cardiac program. Utilizing chromatin immunoprecipitation, next generation sequencing, bioinformatics, predicative biological principals, and or RT-qPCR, we will be able to determine if a correlation exists between the regions of MESP1 binding and Histone3 post-translational modification for activation and repression.

2. Chapter 2 Genome-Wide Identification of MESP1 Targets Demonstrates Primary Regulation Over Mesendoderm Gene Activity.

Published Work

Soibam, B., Benham, A., Kim, J., Weng, K. C., Yang, L., Xu, X., Robertson, M., Azares, A., Cooney, AJ., Schwartz, R.J., Liu, Y. (2015). Genome-Wide Identification of MESP1 Targets Demonstrates Primary Regulation Over Mesendoderm Gene Activity. *STEM CELLS, 33*(11), 3254-3265.

2.1. Introduction

MESP1, a basic helix-loop-helix (bHLH) transcription factor, is long known for its role in cardiovascular development. Recent evidence indicates that MESP1's role in development is broader than currently recognized. It is the first sign of the nascent cardiac mesoderm, but it also marks the appearance of hematopoietic stem cells, head skeletal mesoderm, and endoderm-derived foregut (Kitajima, Takagi et al. 2000, Chan, Shi et al. 2013). Ablation of both MESP1 and MESP2 tandem genes led to the absence of the heart, also accompanied by loss of most anterior structures (Haraguchi, Kitajima et al. 2001). Identifying the early cell lineages regulated by MESP1 is essential for regeneration of the related critical tissues (heart, blood, gut, etc.). This requires identification of MESP1 direct targets in a specific and unbiased fashion. MESP1 interacts with genomic DNA via ebox elements and regulates downstream gene expression (Oginuma, Hirata et al. 2007). Previously, MESP1 gene targets were identified by a candidate gene approach involving induced overexpression of tagged MESP1 in murine embryonic stem cells followed by microarray analysis of the entire mouse embryonic stem cell (mESC) population transcriptome. From this data, a few candidate genes were

identified, and MESP1 chromatin immunoprecipitated (ChIP) DNA was analyzed to validate them. These studies as a first degree approximation of endogenous MESP1 signaling defined MESP1 as the most critical factor for cardiovascular development (Bondue, Lapouge et al. 2008). Even with the advent of deep sequencing technologies, it is still challenging to identify in an unbiased manner the transcriptome and chromatome (collective chromosomal state) prompted by MESP1-fated cells, especially when MESP1 expression is cell type restricted to a group of early progenitor cells and temporally limited to E6–7.5 during embryogenesis (Saga, Miyagawa-Tomita et al. 1999, Saga, Kitajima et al. 2000), and accordingly transiently expressed in a small percentage (3%–5%) of differentiating mES cells (Saga, Kitajima et al. 2000, Wu 2008).

We have taken steps to circumvent this problem by generating a mouse *MESP1*$^{Cre/1}$: *Rosa26*$^{EYFP/+}$ reporter ESC line (referred to herein as UH3 cells). The endogenous MESP1 promoter drives the expression of a knock-in Cre recombinase, which in turn activates the expression of EYFP in the Rosa26 locus. These cells are hence permanently YFP marked. This enables us to follow MESP1 progeny over time even though the endogenous MESP1 gene becomes repressed. We used this system to study the transcriptome and chromatome of cells that were specified only by endogenous MESP1 signaling. Surprisingly, MESP1 primarily directs the appearance of mesendoderm, an early precursor cell, which gives rise to mesoderm (specifying cardiac, blood, and bone cells) and endoderm (specifying foregut endoderm and the pancreas) structures.

2.2. Results

2.2.1. Characterization of an MESP1 Lineage-Tracking ESC Line

An MESP1-lineage reporter mouse ES cell line (UH3) was isolated from E3.5 blastocysts resulted from an *MESP1*$^{Cre/+}$ and *Rosa26*$^{EYFP/EYFP}$ crossing (Figure 2.1A). The cells bore an *MESP1*$^{Cre/+}$*; Rosa26*$^{EYFP/+}$ genotype showed characteristic mES cell morphology (Appendix B Figure S1A) and had a normal male karyotype (Appendix B Figure S1B). To characterize the UH3 cell line and confirm that they undergo differentiation similar to wildtype mES cell lines such as AB2.2, we aggregated the cells in hanging drops, harvested the embryoid bodies over time, and examined the gene expression using quantitative real-time PCR (qRT-PCR). As expected, we observed a transient induction of MESP1 expression in UH3 cells similar to AB2.2 control (Appendix B Figure S2; Figure 2.1B, C). *MESP1* transcript expression coincided with the upregulation of *Gata4* and *Tbx5* genes, while *Nkx2.5* and *Mef2c* expression increased about 1–2 days after MESP1 expression had diminished, marking the induction of the cardiac progenitor cell program (Appendix B Figure S2). Cell surface markers appeared sequentially among which *Cxcr4*, a marker for mesendoderm, appeared concurrently with mesoderm markers *Pdgfra* and *Flk1* (Nelson, Faustino et al. 2008) (Appendix B Figure S2). Increased expression of *Sirpa* and *Alcam*, markers of blood and cardiac myocytes, respectively (Dubois, Craft et al. 2011), followed several days later indicating that the UH3 cell line follows a multi-mesoderm lineage differentiation program (Nelson, Faustino et al. 2008, David, Schwarz et al. 2013).

We further confirmed that UH3 cells undergo cardiac differentiation by well-organized sarcomeres in YFP+ UH3 cells eight days post-hanging drop (Appendix B Figure S1C). Finally, UH3 cells in culture exhibited rhythmic beating (Appendix B Figure S1D).

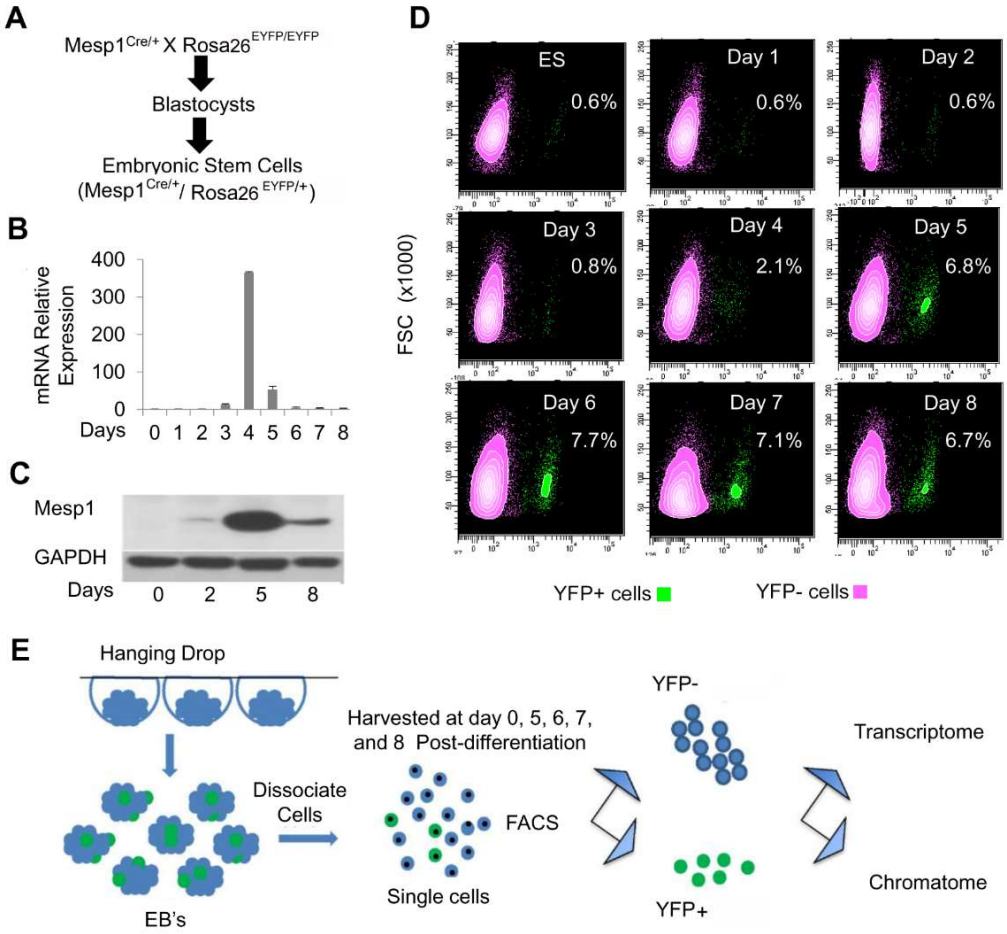


Figure 2.1. Methodology for MESP1-lineage tracing. (A): The strategy in generating mouse *MESP1*$^{Cre/+}$*;Rosa26*$^{EYFP/+}$ reporter embryonic stem cell (ESC) line. (B): The expression of MESP1 mRNA during ESC differentiation, as assayed by realtime RT-PCR. (C): The expression of MESP1 during ESC differentiation, as determined by western blot. (D): Percentage of fluorescence-activated cell sorting-sorted YFP+ MESP1 lineage cells during the course of ESC differentiation. (E): Schematic details of the methodology for MESP1-lineage tracing and whole genome analyses. Abbreviations: EB, * * *; ESC, embryonic stem cell; FACS, fluorescence-activated cell sorting; FSC, forward scatter. I contributed with culturing and FACsorting EB differentiating cells.

2.2.2 Transcriptome of MESP-Marked Progenitor Cells Signified Contribution to the Mesendoderm Lineage

Our strategy, as illustrated by the schematic diagram, was to follow the transcriptome and chromatome of MESP1-fated cells (Figure 2.1E). Briefly, UH3 cells were differentiated by hanging drops, then staged YFP+ cells were isolated to follow MESP1 progeny. Percentages of MESP1-YFP+ cells at each stage are shown in Figure 1D. EYFP signal started to appear as early as day three (Figure 2.1B), while *Mesp1* transcripts peaked at day 4 (Figure 2.1C). MESP1 protein was enriched at day five, as shown by western blot (Figure 2.1C). To identify the transcriptome, RNA was isolated from equal numbers of undifferentiated mES cells and sorted MESP1-YFP+ cells at days 5, 6, 7, and 8 for next-generation sequencing analysis (Done by Benham) (Figure 2.2A). Using mouse refseq genes as the reference annotation (mm9 version), gene expression profiles (Appendix 2 Table S1) were obtained from RNA-Seq data using tophat (Kim, Pertea et al. 2013), and cufflinks (Done by Soibam and Kim) (Trapnell, Williams et al. 2010).

To understand the temporal expression patterns of genes in MESP1-marked progenitor cells, we first identified genes that were upregulated ($p<0.05$) at day five MESP1-YFP+ cells as compared to mES cells using DESeq (Anders, Anders et al. 2010) and performed gene ontology (GO) analyses for this gene set (Done by Soibam and Kim) (Appendix B Table S2, Figure 2B). The highest significance was representatives of mesoderm GO terms such as heart development (89 gene count with p-value of $3.1x10^{-27}$), skeletal system development (91 gene count, p-value of $5.09x10^{-20}$), and vasculature development

(80 gene count with p-value of $9.41x10^{-14}$), and followed by neuron development (58 gene count with p-value of $4.20x10^{-10}$). Currently, there is not a well-defined mesendoderm GO term, but many endoderm GO terms such as respiratory system development (36 gene count, a p-value of $3.33x10^{-7}$), pancreas development (14 gene count, a p-value of $1.57x10^{-4}$) were significantly enriched. Interestingly, upregulated genes at day 6, 7, and 8 MESP1-YFP+ cells, also showed the highest enrichment both for mesoderm and endoderm-associated GO terms (Figure 2.2B). We also noticed a general increase in the number of upregulated genes pertaining to developmental GO terms (Appendix B Table S2) with the progression of differentiation. For instance, the number of heart development associated genes at day 5, 6, 7, and 8 were 89, 85, 110, and 115, respectively. The highly enriched endoderm representative term— respiratory tube development had 32, 34, 43, and 44 upregulated genes in day 5, 6, 7, and 8 YFP+ cells.

To further validate the enriched transcriptome in the YFP+ cells, we asked whether H3K4me3 modification patterns are correlated with the transcriptome quantification. High-quality H3K4me3 ChIP-Seq was performed for ES and day five YFP+ cells (Done by Kim) (Appendix B Figure S3A, S3B). Next, broad peaks were identified using MACS2 (https://github.com/taoliu/ MACS/, an updated version of study in ref. (Zhang, Liu et al. 2008)) with an false discovery rate (FDR) <0.01. H3K4me3 peaks were preferentially located near transcription start sites of genes (Done by Soibam and Kim) (Appendix B Figure S3C, S3D, and Table S3). Genes with H3K4me3 peaks in their promoters had significantly higher expression levels (more than a fivefold mean expression) than those without H3K4me3 peaks

at their promoters (Figure 2.2C). This correlation between the H3K4me3 ChIP peaks and RNA-Seq expression profiles validates the set of actively transcribed genes from our RNA-Seq data, agreeing with observed association of active promoters with H3K4me3 modifications. As representative genes, MESP1 itself was transitorily expressed on day five (Figure 2.2D), agreeing with RT-PCR findings (Figure 2.1C). *Foxa2*, a classic marker of endoderm, was expressed at higher levels in YFP+ marked cells at day six, after the fall of *Mesp1* mRNA (Figure 2.2D). Nkx2-5, a classic cardiac marker, was enriched later in YFP+ cell at the final time points (Figure 2.2D), while expression of Eomes, a mesendoderm marker, correlated to that of MESP1 (Figure 2.2D). Hierarchical clustering of genes associated with heart development, mesoderm development, and endoderm development GO terms revealed subsets of genes that were specific for differentiation stages (Figure 2.2E; Appendix B Figure S4). Next, to determine whether MESP1 YFP+ cells and YFP+ cells showed different expression signatures for mesendoderm lineage, we identified differentially expressed genes using DESeq (Done by Soibam and Kim) ($p<0.05$). Genes upregulated in YFP+ cells were consistently enriched in heart development related terms at day 5, 6, 7, and 8 (Figure 2.2F; Appendix B Table S4). Differentially expressed genes (at day 5, 6, 7, and 8) between YFP+ and YFP- cells showed stronger enrichment of mesoderm and endoderm lineage terms in YFP+ cells (Figure 2.2F; Appendix B Table S4). Thus, our data indicate that the MESP1-marked cells express a unique molecular signature, and are destined to become cardiac progenitors and other

subsequent mesendoderm-derived lineages after the transient appearance of MESP1.

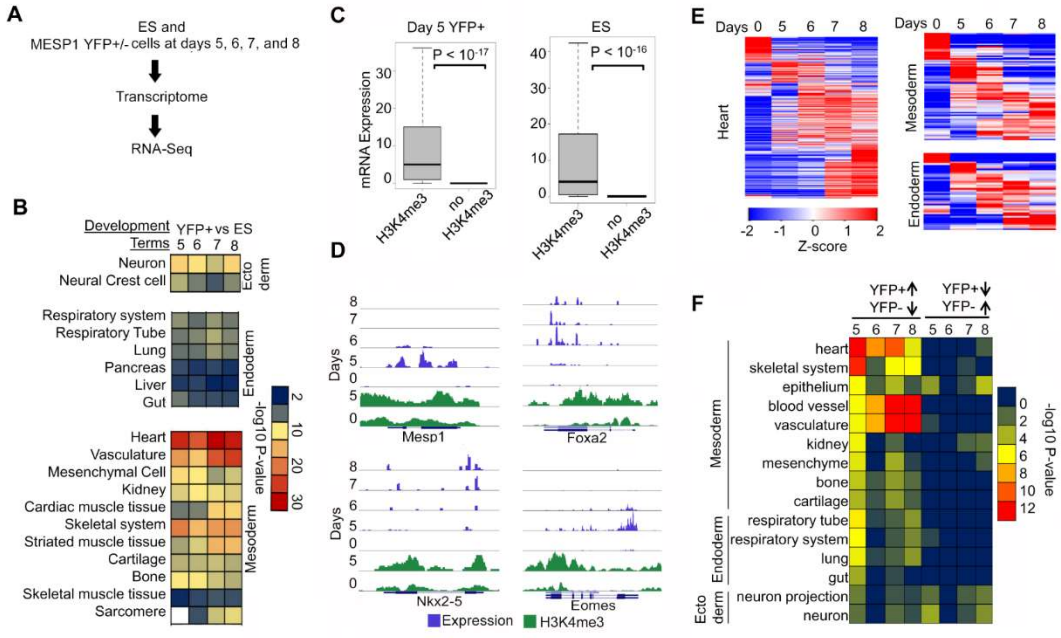


Figure 2.2. Transcriptome of MESP-marked progenitor cells signified contribution to the mesendoderm lineage. (A): Transcriptome profiling of MESP1 YFP+ cells using RNA-Seq. (B): Enriched gene ontology (GO) developmental ontology terms associated with upregulated genes. Comparisons were made at multiple time points (days 5, 6, 7, and 8) between YFP+ and undifferentiated embryonic stem cells. (C): Correlation of H3K4me3 signal with mRNA expression levels. The H3K4me3 signal for a gene was computed as the normalized read count in a 2-kb interval centered at the transcription start site of the gene. (D): Dynamic expression profiles of *Mesp1*, *Foxa2*, *Nkx2-5*, and Eomes in the form of mapped RNA-Seq reads and H3K4me3 ChIP-Seq reads to mouse genome. The scales between the RNA-Seq alignment profiles across different days were adjusted for visual comparison. Similar adjustment was made between embryonic stem cell and day five for H3K4me3 profiles. (E): Hierarchical clustering of genes associated with GO terms—mesoderm development, endoderm development, and heart development. The robust z-score for each gene across the samples is reported. The cosine similarity was used as the distance metric. Detailed heat map at individual gene level is accessible in Appendix B Figure S4. (F): GO terms for differentially expressed genes between YFP+ and YFP- cells at days 5, 6, 7, and 8. Abbreviation: mESC, mouse embryonic stem cell. (I contributed with ChIP-seq library generation and Immunoprecipitation of target protein)

2.2.3 Identification of MESP1 Genomic Binding Sites by ChIP-Sequencing

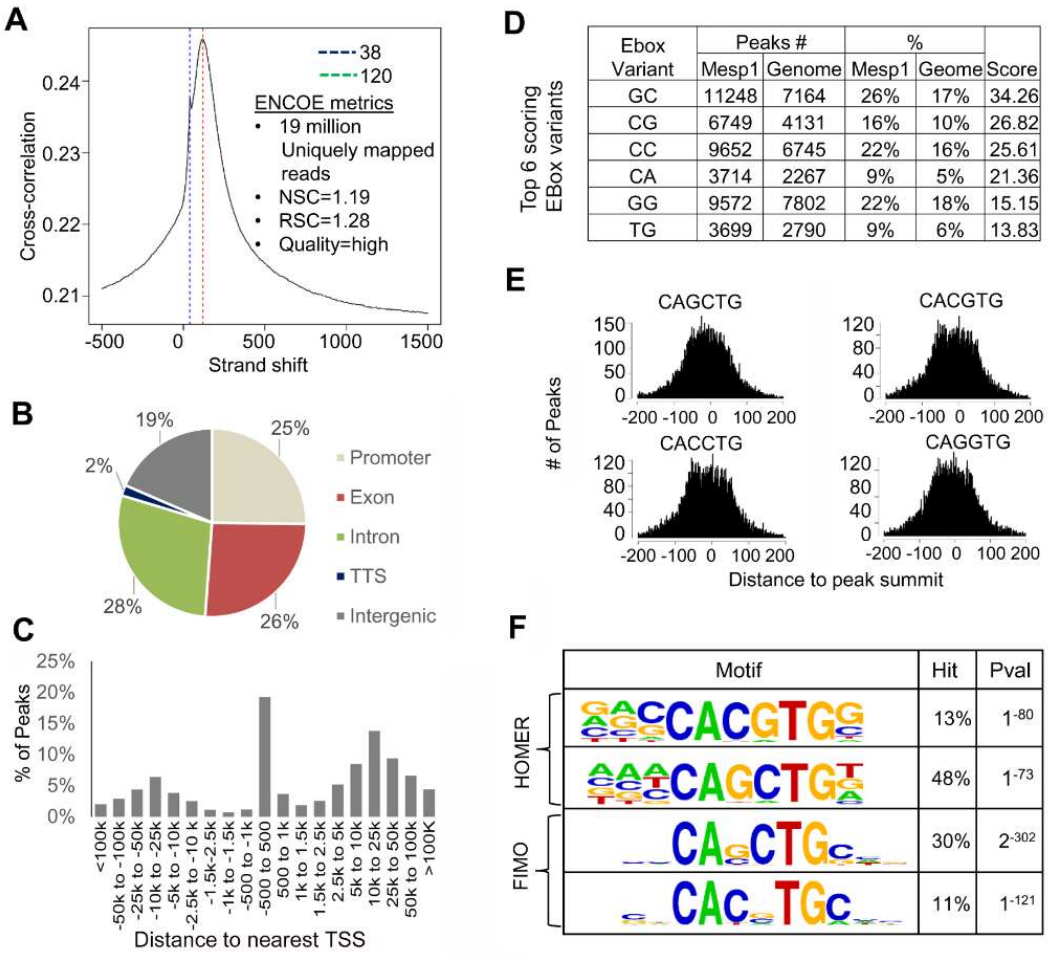
To identify MESP1-binding sites, we performed MESP1-ChIP on day four differentiating ESCs and used next generation sequencing technology to obtain the enriched DNA fragments in the form of raw reads of 40 nt from a single end (Done by Kim, Weng, and Yang). These raw reads were aligned to the mouse genome allowing a maximum of two mismatches per read (Done by Soibam and Kim). The alignment profile of MESP1 ChIP-Seq data passed data quality criteria by ENCODE (Landt, Marinov et al. 2012) (Figure 2.3A, 19 million uniquely mapped reads, relative strand correlation (RSC) 51.27, Qtag51, normalized strand correlation [NSC]51.18). We also generated an appropriate input DNA from the same cells at day four (Done by Kim, Weng, and Yang). We applied widely used MACS2 to the MESP1 ChIP-Seq data and the control background to obtain 43,346 binding sites (peaks) (Appendix B Table S5) with a stringent p-value cutoff of 10^{-8} (Done by Soibam and Kim).

Genomic annotation of the peaks revealed that about 25% of MESP1 peaks resided in promoter (61 kb of nearest transcription start site [TSS]), 26% within the introns, 28% in Exons, and 19% in intergenic regions (Figure 2.3B); further analysis of the positional distribution of peaks from the nearest transcription start site (TSS) showed that majority of the promoter peaks (6,448 out of 9,444) were located within 500 bp from the nearest TSS (Figure 2.3B). Interestingly, 50% of peaks were located at intergenic or intragenic distal enhancers (>10 kb from TSS).

2.2.4 MESP1-Bound Sequence Characteristics

MESP1 directly binds gene regulatory regions containing basic helix-loop-helix binding sites or eboxes (Haraguchi, Kitajima et al. 2001, Bondue, Lapouge et al. 2008). To determine the enriched sequence motifs, we first examined ebox sequences in MESP-bound peaks. A total of 78% peaks contained at least one canonical ebox (CANNTG). Using a logistic regression model for ebox occupancy and adjusted GC content in the peaks (Appendix A Materials and Methods), we found that 12 of 16 variants of ebox sequences had significant enrichment ($p<1x10^{-06}$) in MESP1 bound peaks compared with the random background set (Done by Soibam and Kim) (Appendix B Figure S5). The six highest scoring variants with at least 10% occupancy are shown in Figure 2.3D, and all these six variants passed ENCODE's standard of motif occupancy of at least 10% of peaks. Strongest preference was observed for ebox variant "GC" followed by "CG" (Figure 2.3D). Overall, the variant CASSTG (S stands for C or G) was preferred over the rest of the other variants; 26%, 16%, 22%, and 22% of peaks have the CAGCTG, CACGTG, CACGTC, and CACCTG motif, respectively. The distribution of these four variants on the MESP1 peaks was also clustered at the vicinity of the peak summits (Figure 2.3E). To supplement the ebox searches, we used HOMER (Heinz, Benner et al. 2010) and FIMO (Grant, Bailey et al. 2011) to scan for the presence of known JASPAR (Mathelier, Zhao et al. 2014) binding motifs in MESP1 bound regions. Interestingly, the top enriched motifs by both HOMER and FIMO comprised of different versions of ebox motifs (MYCN, MYOD1, TCF3, TCF12, MYOG, AP4, PFTA, MAX) (Appendix B Figure S6); the top two variants being "CG"

and “GC” (Figure 2.3F). This analysis agrees with our earlier observation that CAGCTG and CACGTG are the top two MESP1 preferred binding motifs. Analysis using HOMER revealed difference in the immediate flanking bases for “CG” and “GC” variant (Figure 2.3F). Similar to the previous study on MYOD (Cao, Yao et al. 2009), MESP1 binding is dependent on flanking and internal nucleotides of CANNTG ebox.



Ebox Variant	Peaks #		%		Score
	Mesp1	Genome	Mesp1	Geome	
GC	11248	7164	26%	17%	34.26
CG	6749	4131	16%	10%	26.82
CC	9652	6745	22%	16%	25.61
CA	3714	2267	9%	5%	21.36
GG	9572	7802	22%	18%	15.15
TG	3699	2790	9%	6%	13.83

Figure 2.3. Genome-wide characterization of MESP1 binding sites. (A): Quality control of MESP1 ChIP-Seq data using the ENCODE standard. Cross-correlation curve peaks approximately at the fragment length 120 bp and not at the read length (40 bp). Other ChIP-Seq quality metrics such as normalized strand coefficient (NSC), relative strand correlation (RSC) were above the standard suggested by ENCODE (according to ENCODE, minimum NSC, and RSC values should be 1.05 and 0.8, respectively). (B): MESP1 peaks annotated as intergenic, promoter (within61 kb of transcription start site), Exon, intron, and TTS. (C): Histogram of number of MESP1 peaks with respect to the nearest transcription start site. MESP1 peaks were found in the promoter as well as in distal intragenic or intergenic regions. (D): The top four variants of ebox motifs enriched in MESP1 peaks compared with background ($p<10^{-16}$). The statistical significant difference was computed using a logistic regression of E-box occupancy and adjusted GC content in the MESP1 peaks (Appendix A Materials and Methods) compared with background sequences. The number of MESP1 peaks and the background sequences containing the ebox variant along with motif enrichment scores are also indicated. (E): Distribution of top four variants of eboxes along MESP1 peaks. The eboxes are clustered around the peak summits. (F): Top Enriched motifs in MESP1 peaks obtained using two different methods (HOMER and FIMO). Both HOMER and FIMO predicted the same variants (“GC” and “CG”) as the two top motifs in MESP1 peaks, which matches the result in (D). Abbreviations: NSC, normalized strand correlation; RSC, relative strand correlation; TSS, transcription start site. (I contributed in analyzing NGS ChIP-seq data)

2.2.5 MESP1 Targets Genes Involved in Mesendoderm Formation

To identify gene targets associated with ChIP-Seq peaks, we compiled a list of 14,006 “potential MESP1 targets” by assigning each peak (43,346 peaks) to the gene whose TSS was closest. We further refined the list by incorporating the MESP1-YFP+ cells transcriptome. We retained 6,470 targets which exhibited differential expression (RNA-Seq, $p<0.05$) at day five YFP+ cells compared with ESCs (Done by Soibam and Kim) (Figure 2.4A). 3,201 and 3,269 were MESP1 activated and repressed targets (Appendix B Table S5; Figure 2.4A), depending on either an increase or decrease in their expression at day five YFP+ cells compared with ESCs. Promoters at MESP1 activated targets showed a significant amount of increase in H3K4me3 modification compared with the repressed targets (Appendix B Figure S7).

To determine the biological functions of MESP1 direct targets, we performed separate GO analysis for MESP1 activating and repressing targets using mouse genome as the background (Done by Soibam and Kim). GO terms pertaining to both mesoderm development ($p=1.89x10^{-9}$) and endoderm development ($p=3.98x10^{-5}$), a variety of signaling pathways (Shh, TGF, BMP, Wnt, Notch, and FGF), and next-tier GO terms denoting mesendoderm-derived organogenesis were significantly enriched in the MESP1 activated targets (Table 2.1; Figure 2.4B). Thus, MESP1 activation targets are strongly enriched for mesoderm and endoderm developmental GO terms, indicating that MESP1 directly regulates developmental pathways pertaining to mesendoderm other than strictly a cardiac lineage. However, the MESP1 repressing targets were not

enriched in mesendoderm related terms, but in GO terms pertaining to broad biological processes such as cell cycle processes (Appendix B Table S5).

To better understand the dynamic expression patterns of MESP1 activated targets, we next performed unsupervised hierarchical clustering of their expression profiles (mES cells and YFP+ cells from days 5, 6, 7, and 8) (Done by Soibam and Kim) (Figure 2.4C). Our analysis revealed nine groups of targets showing unique temporal expression patterns (Figure 2.4C). Cluster assignment for the targets is provided in Appendix B Table. S5. To further understand the functions associated with these clusters of targets, we performed GO analysis for each cluster. Interestingly, targets whose expression transiently peak at day five (cluster 9 in Figure 4C) showed the highest enrichment in gastrulation, pattern specification, and embryonic morphogenesis (Figure 2.4D). Mesendoderm-derived organogenesis was associated with MESP1 targets whose expression became most prominent at day six or beyond (Figure 2.4D). Such patterns indicate that MESP1 target genes form a coordinated gene network that drives mesendoderm formation and subsequently lineage specifications.

We also filtered the “potential MESP1 targets” through differentially expressed genes between day five MESP1 YFP+ and YFP- cells (Done by Soibam and Kim) (Figure 2.4A, right branch). This comparison yielded 475 activated and 796 repressed targets. This set of MESP1 activated targets also showed strong enrichment for mesoderm ($p=1.13x10^{-7}$) and endoderm development ($p=0.0046$), while the repressed targets showed no enrichment for these terms (Figure 2.4B; Appendix B Table S5). GO terms for ectoderm development showed enrichment

in repressed targets (Figure 2.4B) but not in activated targets. This indicates that MESP1 activated targets contribute to mesoderm and endoderm programs. Other GO terms associated with mesendoderm-derived organogenesis also showed enrichment among MESP1 activated targets when referenced to day five YFP- cells (Figure 2.4B). The 797 MESP1 repressed targets showed small enrichment for endoderm-associated terms pertaining to lung development (p=0.0016), pancreas development (p=0.0153). But, these GO terms also showed enrichment in MESP1 activated targets. Terms related to gut development were absent in the repressed targets, which indicates that MESP1 YFP+ cells contribute more to gut lineages than the YFP- population. Using two different reference systems (mES cells and day five YFP- cells), we identified two sets of MESP1 directly activated targets, and both sets showed strong enrichment of both mesoderm and endoderm GO terms. Thus, MESP1 directly regulates developmental pathways pertaining to mesendoderm other than strictly a cardiac lineage.

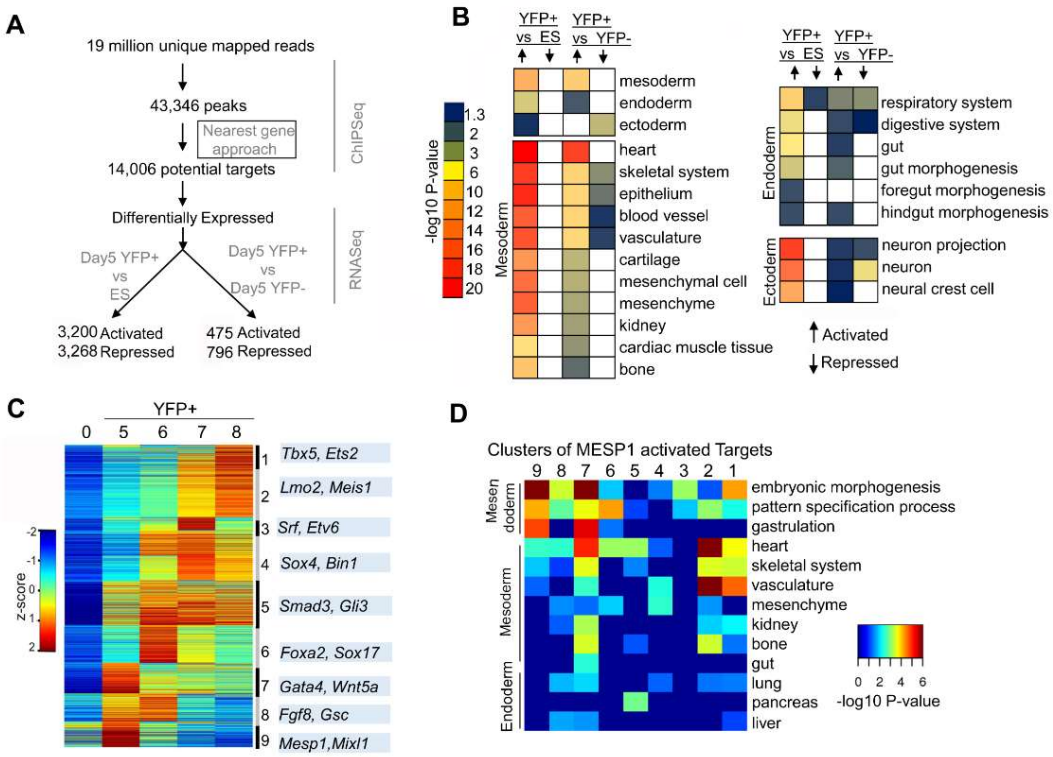


Figure 2.4. Functional assessment of MESP1 targets. (A): Stepwise filtering in identifying high-confidence MESP1 activation targets. First, MESP1 peaks were identified from 19 million unique mapped reads from ChIP-Seq. Each peak was assigned to a gene with its transcription start site nearest to the peak. The genes whch showed differential expression ($p<0.05$) between embryonic stem cell and day five YFP+ cells were retained as MESP1 direct targets, and we obtained 3,200 MESP1 activated, and 3,268 repressed targets, respectively. In a second approach, we compared day five YFP+ cells to YFP+ cells and obtained 476 activated and 797 repressed targets. (B): Gene ontology (GO) analysis of MESP1 activated and repressed targets. The color scale in the heat map is shown as -log10 p value. (C): Clustering of MESP1 activation targets on the basis of their temporal expression pattern. Hierarchical clustering was performed using the fragments per kilobase of exon per million fragments mapped (fpkm) values at days 0, 5, 6, 7, and 8 YFP+ cells. Nine maximally homogeneous clusters were obtained; example genes for each cluster are also indicated. Cluster assignment for the MESP1 activated targets can be assessed in Appendix B Table S5. (D): Individual GO analysis of MESP1 activated genes belonging to the nine clusters. The color scale is indicated in -log10 p value. Abbreviation: mESC, mouse embryonic stem cell.

Table 2.1. Gene Ontology analysis of high-confidence MESP1 activation targets

Germ layer GO terms enriched for MESP1 activation targets

GOID	Term	Count	OR	*p*-Value
GO:0007498	Mesoderm development	30	0.989772	1.90E-09
GO:0007492	Endoderm development	15	0.494886	3.98E-05
GO:0007398	Ectoderm development	31	1.022765	0.023407
GO:0048339	Paraxial mesoderm development	6	0.197954	0.055491

Developmental and signaling GO terms enriched for MESP1 activation targets

Germ layer	GOID	Term	Count	OR	*p*-Value
MES	GO:0007507	Heart development	107	3.530188	1.75E-29
MES	GO:0060429	Epithelium development	101	3.332234	5.28E-18
MES	GO:0001501	Skeletal system development	102	3.365226	8.69E-17
MES	GO:0001944	Vasculature development	91	3.002309	2.04E-15
MES	GO:0001568	Blood vessel development	88	2.903332	1.14E-14
MES	GO:0060485	Mesenchyme development	33	1.08875	1.50E-14
MES	GO:0051216	Cartilage development	38	1.253712	4.28E-11
MES	GO:0001822	Kidney development	46	1.517651	5.89E-11
MES	GO:0014706	Striated muscle tissue development	51	1.682613	9.18E-11
MES	GO:0060348	Bone development	44	1.451666	2.86E-08
MES	GO:0048738	Cardiac muscle tissue development	25	0.82481	1.70E-06
MES	GO:0048534	Haematopoietic or lymphoid organ development	61	2.012537	0.006996
EN	GO:0060541	Respiratory system development	44	1.451666	1.48E-07
EN	GO:0030324	Lung development	39	1.286704	1.11E-06
EN	GO:0048565	Gut development	19	0.626856	7.08E-06
EN	GO:0031016	Pancreas development	18	0.593863	1.37E-05
EN	GO:0055123	Digestive system development	17	0.560871	1.69E-05
EN	GO:0001889	Liver development	18	0.593863	1.41E-04
EC	GO:0031175	Neuron projection development	85	2.804355	1.24E-16
EC	GO:0048666	Neuron development	97	3.200264	1.55E-13
EC	GO:0014032	Neural crest cell development	22	0.725833	6.27E-10
SIG	GO:0016055	Wnt receptor signaling pathway	61	2.012537	1.60E-16
SIG	GO:0030509	BMP signaling pathway	17	0.560871	2.72E-08
SIG	GO:0007219	Notch signaling pathway	25	0.82481	5.12E-07
SIG	GO:0007179	Transforming growth factor beta receptor signaling pathway	23	0.758825	7.96E-07
SIG	GO:0007224	Smoothened signaling pathway	13	0.428901	5.66E-04

Count: number of genes associated with the GO team.

Abbreviations: EN, endoderm; EC, ectoderm; MES, mesoderm; ORs, the odds ratio between the numbers of predicted genes and observed genes in the GO team; SIG, signaling pathways.

Since “mesendoderm development” is only partially completed, there is not a proper GO term (Ashburner, Ball et al. 2000), to test if MESP1 targets key mesendoderm regulators. We investigated individual genes that are directly involved in mesendoderm development. Mesendoderm markers such as *Gata4*, *Eomes*, *Wnt5a*, *Wnt5b*, *Mixl1*, *T*, *Gsc*, and *Wnt3* were among MESP1 activated targets (Figure 2.5A). We found that MESP1 may regulate itself, perhaps through a binding site located about 28 kb from the TSS (Figure 2.6E). Hematopoietic transcription factors such as *Tal1*, *Meis1*, and *Lmo2* were also direct MESP1 activated gene targets, while other key regulatory blood factors, *Gata1*, *Hbby*, and *EBf1* and cardiac progenitor markers, *Nkx2-5*, *Mef2c*, and *Tnnt2* were not (Appendix B Table S5; Figure 2.5C). In YFP+ cells, MESP1’s temporal expression pattern correlated well with the appearance of mesendoderm gene expression (such as *Gata4*, *Mixl1*, *Wnt5b*, *Wnt3*, and *T*) while significantly differing from that of key cardiac and blood genes (Figure 2.5C). Thus, MESP1 indirectly regulates the cardiac or hematopoietic differentiation programs in a discrete MESP1 enriched population (Bondue and Blanpain 2010). We also examined the expression of these key mesendoderm markers in the YFP- population. Most of the mesendoderm markers except T showed significantly higher expression at day five in YFP+ cells than YFP- cells (Figure 2.5A). MESP1 activated targets, which are involved in cardiac and hematopoietic programs, are higher in YFP+ cells compared with YFP- cells at later stages of differentiation (after day five), indicating an indirect regulation of these programs by MESP1 (Figure 2.5A).

Previous studies using MESP1 overexpression followed by ChIP-PCR reported that MESP1 bound to conserved E-box sites at regions close to the TSS of *Foxa2* (5 kb), *Sox17* (4 kb), *Gsc* (4.5 kb), and T (1.5 kb), lead to rapid down-regulation of these genes (Bondue, Lapouge et al. 2008, Bondue and Blanpain 2010). In our unbiased approach, endoderm markers *Foxa2* and *Sox17* were MESP1 activated targets (Figure 5A). It is possible that upon forced induction, MESP1 binds to regions indicated by the previous study (Saga, Kitajima et al. 2000) close to TSS and down-regulates the important endoderm markers, likely through an inhibitory complex. In YFP+ cells, endoderm genes *Foxa2*, *Cer1*, and *Sox17* began to appear at day five and peak at day six, while MESP1 levels fell and disappeared. These data suggest that a subset of MESP1 YFP+ cells, at a later stage showing definitive endoderm (DE) lineage characteristics did not depend on continuous MESP1 gene expression. *Sox17*, *Foxa2*, and *Cer1* were expressed at higher levels in YFP- cells at days five, six, and eight (only Foxa2) indicating that YFP- cells primarily contribute to the endoderm lineage. To prove the concept, we costained YFP+ cells with T, Foxa2, or Sox17 in day 5 differentiated cells (Figure 2.5B). Both T and Foxa2 were highly prevalent, agreeing with efficient mesendoderm formation. YFP signals were located in the cytosol and were frequently detected in T or Foxa2 positive cells. Sox17 signals were more restricted in a small number of clustered cells, agreeing with limited endoderm differentiation. Co-staining of Sox17 and YFP was also evident. Although T, Foxa2, and Sox17 were more prevalent in YFP- cells, our data suggest that MESP1 is associated with at least a population of mesendoderm cells which

contribute to endoderm lineages, but only after the reduction of MESP1 gene activity.

We further explored GO terms associated with MESP1 activated targets which were differentially expressed between YFP+ and YFP- cells at different time points during differentiation. For this, we compiled two lists of MESP1-activated targets. The first group consists of targets which were strictly expressed at higher levels in YFP+ cells during differentiation, while the second group was strictly expressed at the higher expression in YFP- cells during differentiation. The first group showed more enrichment in mesoderm and endoderm-derived organogenesis, the majority being in the endoderm. (Appendix B Figure S8). However, some of the endoderm terms such as respiratory and lung development were also enriched in the second group. This indicates that some of the genes, which are activated post-MESP1 expression in the YFP+ cells, are also expressed in the YFP- population and may contribute to endoderm lineages. Such genes are most likely regulated by other factors, independent of MESP1, in the YFP- population.

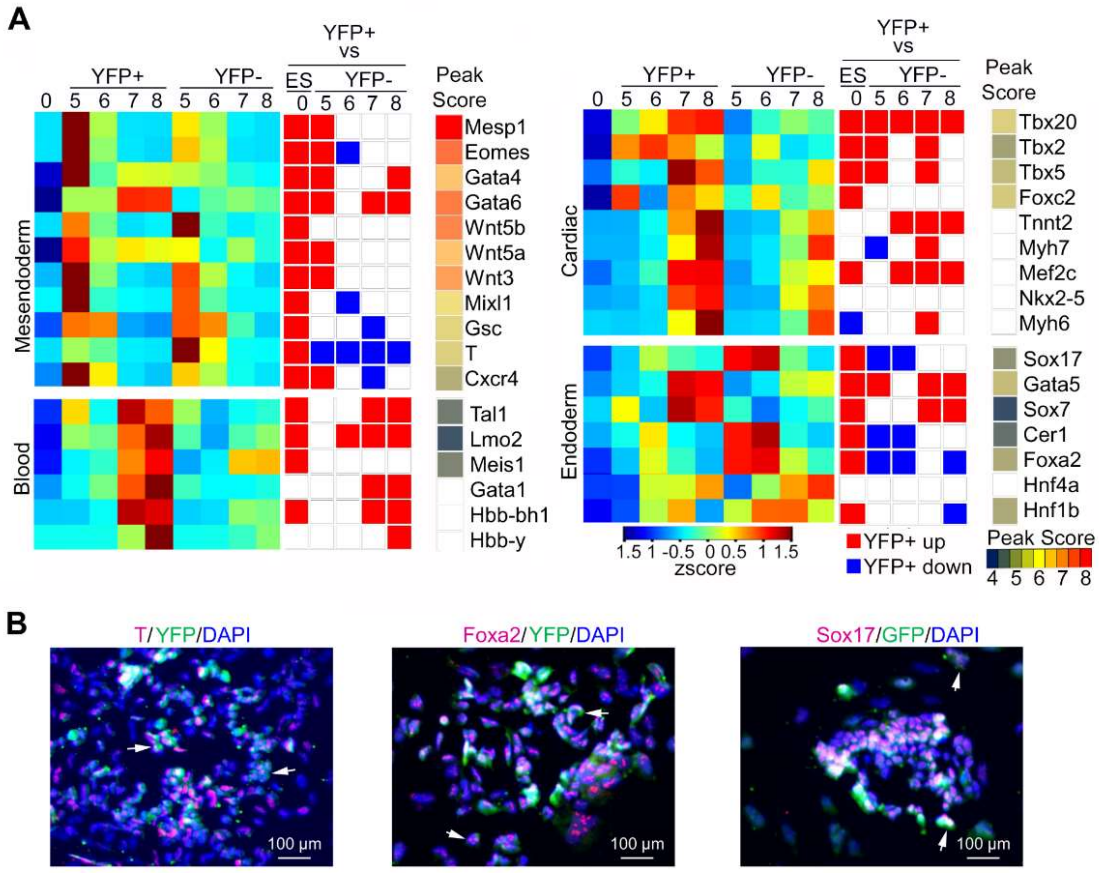


Figure 2.5. MESP1 directly regulates mesendoderm genes. (A): mRNA expression levels, differential expression between YFP+ and YFP- cells, and ChIP-Seq peak scores of key mesendoderm, endoderm, cardiac, and blood genes. For each lineage-specific group of genes, there are three heat maps placed in three subpanels. The first heat map represents expression profiles of the key markers in mouse embryonic stem cell (mES), YFP+ and YFP- cells at different time points. The middle heat map indicates if the gene was differentially expressed between YFP+ versus mESCs/YFP cells at different time points. In this heat map, "red," "blue," and "white" colors indicate up-regulation in YFP+, down regulation in YFP-, and not differentially expressed, respectively. The third map represents the peak score from the MESP1 ChIP-Seq associated with the gene. The color scale of peak score is shown as log2 of read-count. In case of multiple peaks assigned to the same gene, the one with the highest score was used. The peak scores for non-MESP1 activated targets are shown as blank. (B): Co-immunostaining analysis of MESP1-lineage and T, Foxa2, and Sox17. The arrows point to cells that show nuclear T, Foxa2, or Sox17 staining and cytosol YFP staining. Abbreviations: DAPI, 40,6-diamidino-2-phenylindole; mESC, mouse embryonic stem cell; GFP, green fluorescent protein.

2.2.6. MESP1 May Recruit H3K27ac Epigenetic Modification to Distal Regulatory Regions

To assess if MESP1 binding sites might modify local chromatin structure, we compared MESP1 ChIP-Seq data to H3K27ac ChIP-Seq data from enriched mesoderm cells (Wamstad, Alexander et al. 2012), and also to H3K4me3 tracks generated in this study (Done by Soibam and Kim). In the enriched 43,346 MESP1 binding sites, the amount of H3K27ac showed a strong correlation (Pearson correlation coefficient 50.91, Figure 2.6A) with a MESP1-binding signal. A similar correlation was also observed (Pearson correlation coefficient 50.84, Figure 2.6B) in intergenic or intragenic distal peaks (>10 kb from TSS).To determine if H3K27ac at MESP1 bound regions are MESP1-binding dependent, we compared the amount of H3K27ac at MESP1-bound regions in mesoderm cells to that in ESCs (Done by Soibam and Kim). An increase in the amount of H3K27ac was observed between ESCs and mesoderm cells at MESP1-bound peaks and also only at distal enhancers (Appendix B Figure S9; Figure 2.6C, 6D). Core mesendoderm genes, *Gata4*, *Wnt3*, *Gsc*, *Gata4*, *Gata6*, and *Mixl1*, displayed enhanced H3K27ac histone modification in MESP1-binding regions and heightened H3K4me3 epigenetic modifications at their promoters in comparison with virtually nil levels in ESCs (Figure 2.6). MESP1 DNA binding displayed a de novo appearance of H3K27ac modification at endogenous binding regions of MESP1, but H3K4me3 levels were already present in replicating mES cells (Figure 2.6E), likely due to activation by Oct4 and Lef1/b-catenin (Li, Yu et al. 2013). We propose that MESP1 target sites

that incorporate the genetic signature of increased H3K27ac and H3K4me3 are likely participants in the acquisition of more differentiated states and constitute an essential mechanism for the transition of pluripotency to mesendoderm.

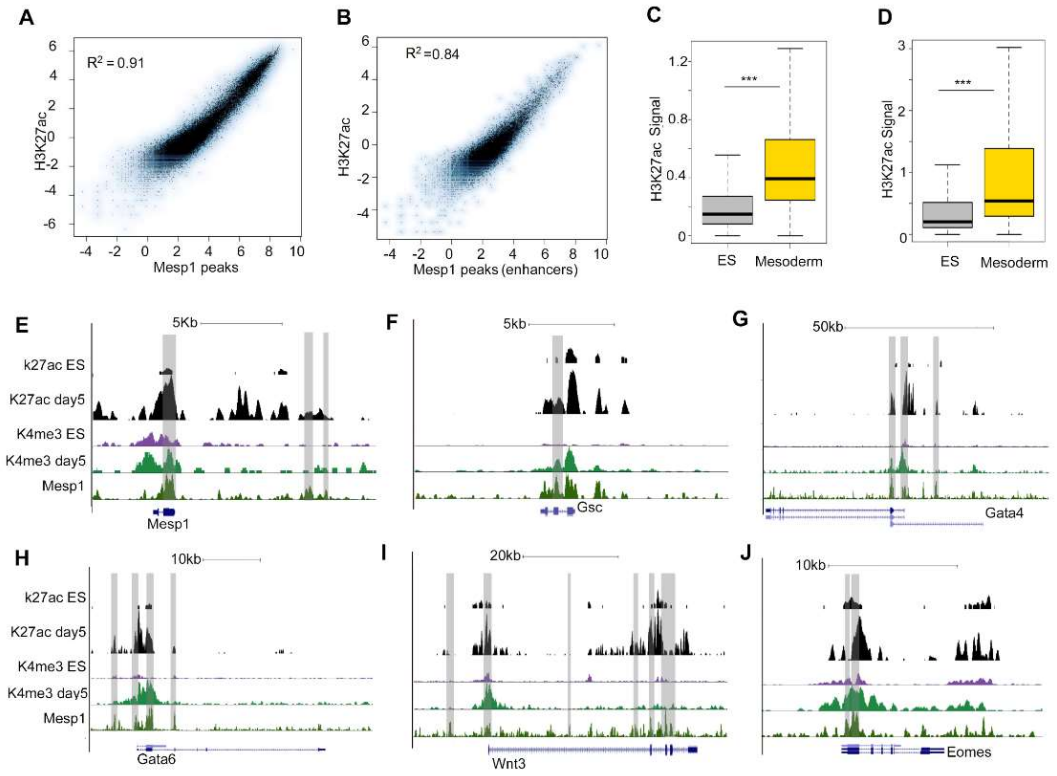


Figure 2.6. MESP1-binding correlates globally with H3K27ac epigenetic modification. (A): H3K27ac signal correlates globally with MESP1 ChIP signals. At all the MESP1 peak locations, the H3K27ac ChIP signal correlated with MESP1 ChIP signal with a Pearson correlation of 0.91. (B): H3K27ac signal correlates with MESP1 ChIP Signal at MESP1 enhancer peaks. At all the MESP1 peak enhancer locations, the H3K27ac ChIP signal in mesoderm cells correlated with MESP1 ChIP signal with a Pearson correlation of 0.84. (C): Increase in H3K27ac signal at MESP1 peak regions from mouse embryonic stem cell (mESC) to mesoderm cells. (D): Increase in H3K27ac signal at MESP1 peak enhancer regions from mES cells to mesoderm cells. (E–J): Alignment profiles of ChIP-Seq reads at loci of selected MESP1 mesendoderm targets. The alignment profiles represent the read depth at resolution of 1 nucleotide. Profiles are shown for MESP1, Gsc, Gata4, Gata6, Wnt3, and Eomes. The identified MESP1 binding regions which are enriched compared with background ChIP-Seq data with $p<10^{-8}$ are shaded in gray. The scales for H3K27ac plots between mES cells and day five were adjusted for visual comparison between the two conditions. Similar adjustment was done for H3K4me3 plots between mES cells and day five. Abbreviation: mES, mouse embryonic stem cell. (I contributed in analyzing NGS ChIP-seq data)

2.3. Discussion

Mesendoderm, an ancient germ layer from worms to frogs, gives rise to both endoderm organs such as liver, foregut, and pancreas and mesoderm organs such as heart, blood, and bone (Rodaway and Patient 2001). Mesendoderm is implicated as the major source of cardiac mesoderm and anterior endoderm in mammals, a part of which can eventually differentiate into hepatocytes and pancreatic cells (Wells and Melton 1999). From *C.elegans* up to *Xenopus*, NODAL, WNT5A/B, PITX2, GSC, MIXL1, EOMES, and GATA4/6 have been shown to participate in mesendoderm induction (Wells and Melton 1999, van den Ameele, Tiberi et al. 2012). The origin of mammalian mesendoderm has not been as well studied, but mesoderm and endoderm cell–cell interactions are especially important. Our model depicts how MESP1 directs a mesendoderm bi-potential developmental pathway and serves as a novel paradigm shift for MESP's role in the cardiopoiesis program. We profiled the transcriptome of a pure population of MESP1-marked cells along with determining the chromatome of endogenous MESP1-bound DNA targets. Surprisingly, MESP1 primarily directs the appearance of mesendoderm instead of the cardiac program per se. Critical mesendoderm modulators including *Mixl1*, *Pitx2*, *Gata4*, *Gata6*, *Wnt5a*, *Wnt5b*, *Sox17*, and *Foxa2* were enriched in MESP1-YFP+ cells before the appearance of cardiac progenitors and myocytes.

Previously, we demonstrated that an endoderm-associated Sry-box transcription factor, SOX17, was essential for cardiac specification in differentiating mES cells acting at least in part via cell-non-autonomous

mechanisms (Liu, Asakura et al. 2007). Recently, we showed in unbiased genome-wide testing, SOX17 expression in ESCs was a prerequisite for the induction of highly diverse cardiogenic transcription factors and cardiac structural genes (Liu, Kaneda et al. 2014). HHEX and CER1 are indispensable components of the SOX17 pathway for cardiopoiesis in mES cells, acting at a stage downstream from MESP1/2. Our demonstration that *Sox17*, *Foxa2*, and *Cer1* are MESP1 targets and their transcripts become enriched in MESP1-YFP+ cells supports the idea that mesendoderm cells initiate a potent auto-regulatory program to direct the earliest cardiac and endoderm progenitors.

We used a stepwise filtration to identify “MESP1-activation targets”. MESP1 ChIP-Seq alone, however, resulted in a total of 43,346 MESP1 binding sites among which most resided both on regions distal as well as close to TSS. The peaks were significant with a stringent p-value of 10^{-8} compared with the background, our ChIP-Seq data passed high-quality criteria recommended by ENCODE, and correlated with H3K27 acetylation indicating that the peaks are specific. Interestingly, ChIP with another lineage-specifying bHLH factor, MyoD, in myoblasts and myotubes yielded a lot of reads in distal enhancer regions, which also showed increased H4 acetylation (Cao, Yao et al. 2009). Why these regulatory regions harbor activating histone modification without net downstream activation is intriguing. Perhaps the bHLH-recruited histone acetylation leads the genes to a “poised” status, ready for expression in the next stage of cellular differentiation, which is an interesting direction to pursue in future studies.

Forced expression studies are held with some degree of skepticism when their activity is extrapolated to the behavior of the endogenous factor. Recently, Tapscott and colleagues provided a litany of cogent reasons why the legitimacy of forced expression studies may be questionable, such as nonphysiological conditions and non- specific DNA binding; even though as a point of fact they did not detect significant differences between their unbiased MyoD ChIP assays and overexpression data (Yao, Fong et al. 2013). We compiled a list of DNA regions which were previously tested for MESP1 enrichment using ChIP-PCR by forced expression (Bondue, Lapouge et al. 2008). About 25% of the MESP1 enriched regions identified were detected by our MESP1 ChIP-Seq. We also compared the genome-wide binding of endogenous MESP1 against the microarray analysis following MESP1-overexpression (Table 1 of the published work by the Blanpain group) (Bondue, Tännler et al. 2011). 155 (145 activated and 10 repressed targets in our data) out of 216 potential MESP1 DNA-binding targets were confirmed (Appendix B Table S6). MESP1 was claimed to promote the expression of cardiac structural genes, such as *Myh7 (b-MHC), Myh6 (a-MHC), Myl1 (MLC1f), Myl2 (MLC2v),* and *Tnnt2 (cTnT)*, but none of these contractile proteins, which appear almost 2–3 days after the transient induction of MESP1, were identified to be direct targets. We also found concordance between the published findings by Lindsley et al, 2007. and ours in which their transient MESP1 expression in mES cells markedly increased the frequency of PDGFRa1 and FLK11 cells (Lindsley, Gill et al. 2007). We noted that *Pdgfra* and *Flk1* are MESP1 targets. Also, MESP1 robustly induces transcription factors that regulate EMT, such as *Snai1* and *Twist*

(Carver, Jiang et al. 2001, Thiery and Sleeman 2006) which were also confirmed by our study, as MESP1 gene targets (Appendix B Table S5). Furthermore, MESP1 failed to induce paraxial mesoderm genes, such as *Meox1, Tcf15, Tbx6,* or *Pax1* (Sakurai, Era et al. 2006) or skeletal, myogenic transcription factors Myod, Myogenin, or Myf5. Besides DE markers such as *Foxa2, Cer1,* and *Sox17,* MESP1 also binds to DNA regions associated with visceral endoderm (VE) markers such as *Sox7* and *Gata5* (Figure 5A). VE cells have been shown to get integrated into definitive endoderm (DE) rather than being displaced by it at a later stage. The expressions of DE and VE markers are more prominent after the disappearance of MESP1 in YFP+ cells (Figure 5A), and the DE markers are more enriched in YFP- cells. These data suggest that MESP1 YFP+ cells contribute to both the VE and DE program; but after the disappearance of MESP1.

We and others previously reported that *Mesp1* transcription is under the regulation of canonical Wnt and T-box factors (T and Eomes) (Costello, Pimeisl et al. 2011, David, Jarsch et al. 2011, Li, Yu et al. 2013). New data from this study suggest that MESP1 regulates the expression of *Eomes*, *Wnt3a*, and *Mesp1* itself. Together MESP1 and these early factors form a self-regulatory network which drives the formation of mesendoderm or a set of it. The regulation of epigenetic marks, MESP1 further specify downstream cardiac genes, though at a later stage when MESP1 itself is no longer expressed. Our study using MESP1-YFP+ cells reveals that MESP1 expression correlates with key mesendoderm markers. First heart-field markers such as *Fgf8, Meis2,* and *Tbx5* and the second heart field markers such as *Smarcd3, Hoxb2, Hoxa1,* and *Cited1* (Lescroart, Chabab et al.

2014) have the highest expression after day five. These observations indicate that MESP1-YFP+ cells are not primarily restricted to first heart field (FHF) and second heart field (SHF), but can give rise to another mesoderm lineage including hematopoietic and endoderm lineage.

To achieve long-term cardiac cell therapy, it's important to understand the regulatory network of cardiac progenitors. In particular, our group found that MESP1 and ETS2 were able to transdifferentiate human dermal fibroblasts into cardiac progenitors (Islas, Liu et al. 2012). Other combinations of transcription factor that successfully converted human fibroblasts to cardiomyocyte-like cells also included MESP1 as one of the transcription factors (Fu, Stone et al. 2013). This study demonstrates that MESP1 primarily drives the activation of genes involved in mesendoderm lineages, which potentially promotes interaction between different germ layers (mesoderm and endoderm) leading to cardiac, blood, and skeletal muscle differentiation. We believe that further elucidation of the genetic makeup of the MESP1-dependent regulatory elements would provide novel strategies in driving mesendoderm in tissue regeneration.

3. Chapter 3 MESP1 Represses the Neuroectoderm Program by recruiting the PRC2 complex to a specific Ebox Variant

3.1. Introduction

MESP1 is a transiently expressed only during a short period of time, from 6.5-7.5 dpc in the posterior region in embryogenesis (Saga, Hata et al. 1996). MESP1-Cre recombinase marked the early progenitors of the mouse cardiac mesoderm lineage even after its down regulation (Saga, Miyagawa-Tomita et al. 1999, Saga, Kitajima et al. 2000, Soibam, Benham et al. 2015, Chiapparo, Lin et al. 2016). Key cardiac gene were still activated within lateral mesoderm tissue but migration anterolateral fails to initiate or is delayed. This loss of function causes embryonic lethality through cardia Bifida and the failure of fusion of the paired heart fields. MESP1-fated cells remain in the posterior region causing malformation of the paraxial mesoderm affecting the formation of posterior tissue development (Saga, Miyagawa-Tomita et al. 1999, Saga, Kitajima et al. 2000). Furthermore, simultaneous disruption of both *Mesp1* and *Mesp2* genes (dKO) led to the complete loss of posterior structures including heart, somites, and gut. Chimera analysis, however, showed that *Mesp1* and *Mesp2* dKO cells still contributed to the formation of somites and gut, but not the heart. These data indicate that Mesp1 and Mesp2 are essential for the formation of the cardiac lineages.

The proper formation of the lateral mesoderm that differentiates into cardiomyocytes, require MESP1, but the initiation of the cardiac program appears to be independent and can be compensated by other factors (Saga, Miyagawa-

Tomita et al. 1999). This suggests that MESP1 may have a more important role in epithelial-mesenchymal transition (EMT) or a potential dominant role of repression of non-mesoderm lineages. Evidence of an MESP1-repressive role can be observed throughout publications but hasn't been completely investigated. However, it has been suggested that MESP1 represses the endoderm and the neuroectoderm program during gastrulation (Lindsley, Gill et al. 2007, Bondue, Lapouge et al. 2008, Soibam, Benham et al. 2015, Chiapparo, Lin et al. 2016). Our recent genomic studies have suggested that MESP1 may be involved in the formation of the mesendoderm germ layer in higher vertebrates by activation of *Gata4, Eomes, Wnt5a, Wnt5b, Mixl1, T, Gsc,* and *Wnt3,* and the hematopoietic transcription factors, *Tal1, Meis1,* and *Lmo2*, challenging the preconception role. (Soibam, Benham et al. 2015). It was also shown that MESP1 does not directly induce the cardiac structural genes, *Myh7 (b-MHC), Myh6 (a-MHC), Myl1 (MLC1f), Myl2 (MLC2v), and Tnnt2 (cTnT)* (Soibam, Benham et al. 2015). Co-staining of T, Foxa2, and Sox17 with YFP show that MESP1 is associated with at least a portion of mesendoderm cells (Soibam, Benham et al. 2015). Transcriptome analysis of FACS purified endogenously activated MESP1 cells showed more repressed genes in comparison with activated genes (Lindsley, Gill et al. 2007, Soibam, Benham et al. 2015, Chiapparo, Lin et al. 2016). We showed that MESP1 primed the cardiac-specified genes through histone modification of target sites probably for later activation by downstream factors (Soibam, Benham et al. 2015).

Here we investigate which developmental programs and gene targets are repressed by MESP1 during mesoderm formation and the mechanism(s) by which

MESP1 can exert repression. Utilizing bioinformatics and publicly available NGS data, I found that MESP1 targets key neuro-ectoderm GO developmental terms and surprisingly has a higher collection of nervous system gene targets than mesoderm developmental gene targets. Computational analysis of MESP1-enriched regions indicated a bias over the two Class I ebox variant in term of function and the recruiting of EZH2 and the deposit H3K27me3. Using induced gain-of-function assay, I found that MESP1 repressed key neuro-ectoderm factors in mouse embryonic stem cells

3.2. Results

3.2.1. MESP1 binding enriches adjacent to multiple neuro-ectoderm regulators

To further elucidate MESP1's repressive function during Mesendoderm formation, I analyzed publically available MESP1 chip-seq raw data from day four EB-differentiated mES cells. MESP1-binding regions were identified by using given peaks size, minimum distance between peaks of 50 bps, filtering based on local fold change of 0 and Poisson p-value of 0.0001 and target genes by assigning each peak with the closest transcription start site (TSS) using HOMER software from the Salk institute (Heinz, Benner et al. 2010). We identified 217,565 binding regions adjacent to 20,005 different TSS. Next, I determined the impact of MESP1 on the total developmental process. I analyzed the gene list through DAVID online GO ontology analysis (Huang, Sherman et al. 2008, Huang, Sherman et al. 2009). I focused primarily on MESP1's effects on developmental processes, by filtering the list for "development." I discovered that a list of MESP1 gene targets has a

potentially large influence on the neuroectoderm developmental lineages, which was actually surprising considering the nature of MESP1 is involved in the development of the heart, head mesenchyme, and paraxial mesoderm formation (Figure 3.1) (Saga, Kitajima et al. 2000).

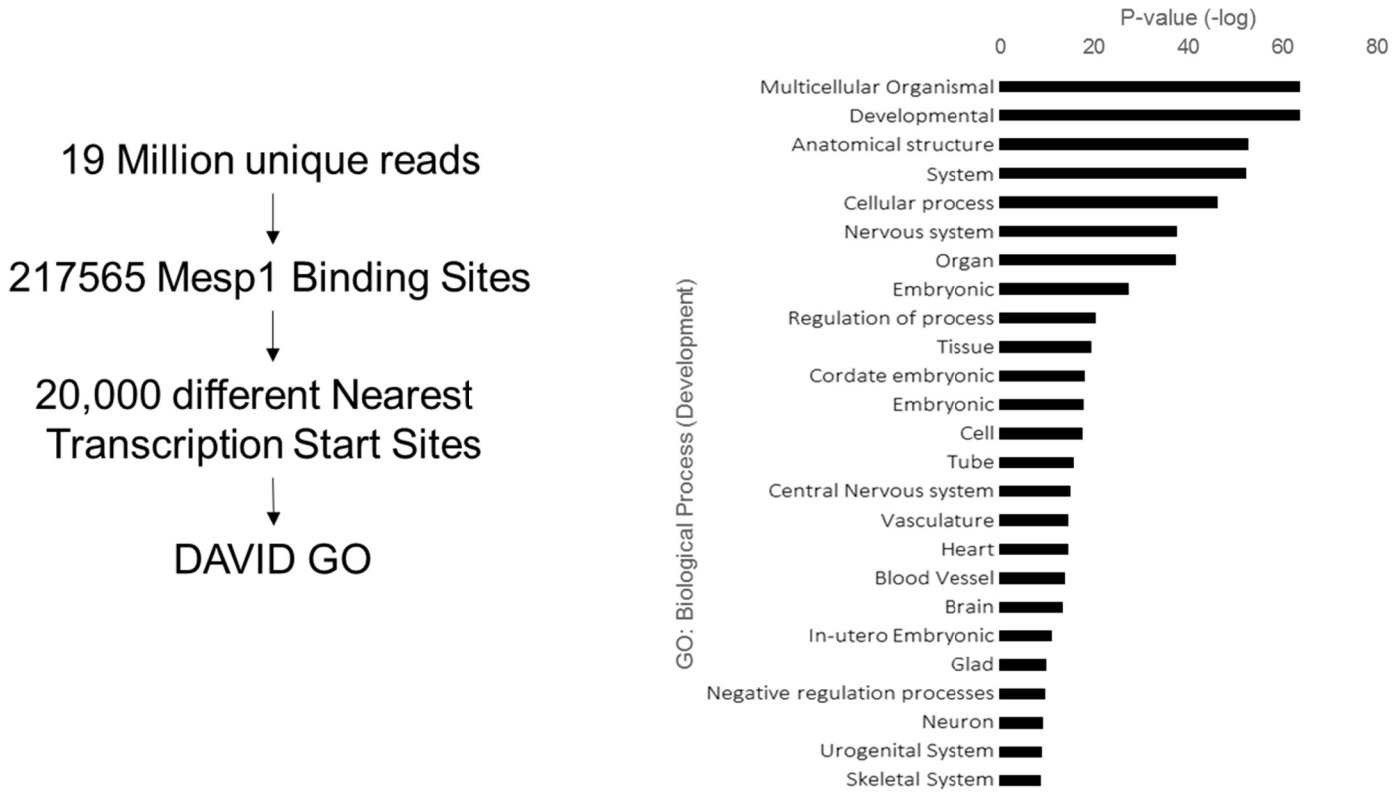


Figure 3.1. Analysis of MESP1 ChIP-seq data. A) Flow-chart of MESP1 ChIP-seq data-analysis. B) GO analysis of all development terms that MESP1 targets in order of significance. Showing by –log(p-value) the degree of significance of each term.

My analysis revealed enrichment for genes involved in the nervous system developmental hierarchy tree over mesendoderm terms. These terms including the nervous system (802 gene count and p-value $4.42x10^{-38}$), central nervous system (345 gene count and p-value $1.39x10^{-15}$), brain (272 gene count and p-value $5.54x10^{-14}$), and neuron (268 gene count and p-value $8.65x10^{-10}$) (Appendix C).

I combed through the genes with each GO term to determine if there were major developmental regulators. My observation revealed that many genes defined by GO terms associated with nervous system were MESP1 targets largely associated with neurogenic differentiation factors *Neurod1, Neurod2, Neurod4, Neurod6* (Lee 1997, Franklin, Kao et al. 2001, Cho and Tsai 2004, Gao, Ure et al. 2009, Ohtsuka and Kageyama 2010), Neurogenin, *Neurog1, Neurog2, Neurog3*, (Ma, Kintner et al. 1996, Lee 1997, Ma, Fode et al. 1999, Kele, Simplicio et al. 2006, Seo, Lim et al. 2007, Lim and Kroll 2008, Ma, Yan et al. 2009, Ohtsuka and Kageyama 2010), and SRY-box containing genes *Sox1, Sox2, Sox3, Sox5, Sox8, Sox10,* and *Sox11* (Wood 2002, Kan, Israsena et al. 2004, Puligilla, Dabdoub et al. 2010, Wegner 2011, Guo, Liu et al. 2014). These data strongly suggests that in addition to mesoderm-specific genes, MESP1 may directly regulate the neuroectoderm developmental gene pathway.

3.2.2. The neuro-ectoderm program is repressed in mesendoderm cells

Since virtually none of the key neuro-ectodermal MESP1 targets were expected to be activated in mesendoderm germ layer, I hypothesized, that MESP1 can potentially act as a repressor on neuro-ectoderm gene targets. First, I wanted

to test if any of the neuro-ectoderm targets of MESP1 were expressed and or repressed within MESP1-progenitor cells. To investigate, I cross-compared all of MESP1-targets with RNA-sequence data representing MESP1 enriched populations following hanging drops of ES cells to form embryoid bodies (EB) on the 5^{th}, and 7^{th} days. Culture conditions were described in our previous publication (Soibam, Benham et al. 2015). I identified and performed separate GO analysis on MESP1-activated (greater than 1.5 fold) and-repressed (less than 0.5 fold) targets to determine the impact that MESP1 had on developmental programs (Figure 3.2). My analysis resulted in six clusters (A-F) with differential patterning determined by comparing the –log of the GO term p-value. The list of GO terms in each cluster can be found in Appendix B, Figure S1.

MESP1 appeared to be a neuro-ectoderm-specific repressor as seen by GO terms based on p-value significance which includes nervous system (111 gene count and p-value $4.66x10^{-11}$), central nervous system (46 gene count and $1.11x10^{-4}$), brain (39 gene count and $7.30x10^{-5}$), hindbrain (16 gene count and p-value $3.65x10^{-5}$), and neuron development (40 gene count and p-value $4.50x10^{-5}$) (Figure 3.2 and Table 1). To measure the gain or loss of function, I analyzed genes involved in nervous system development by its GO term and found key neuro-specific factors involved in development, *Neurod1, Neurog1, Neurog2, Neurog3, Shh, Sox2,* and *Sox3* (Table 3.2). This comparison helped to shed light on MESP1 role, as a direct repressor of neuro-ectoderm processes.

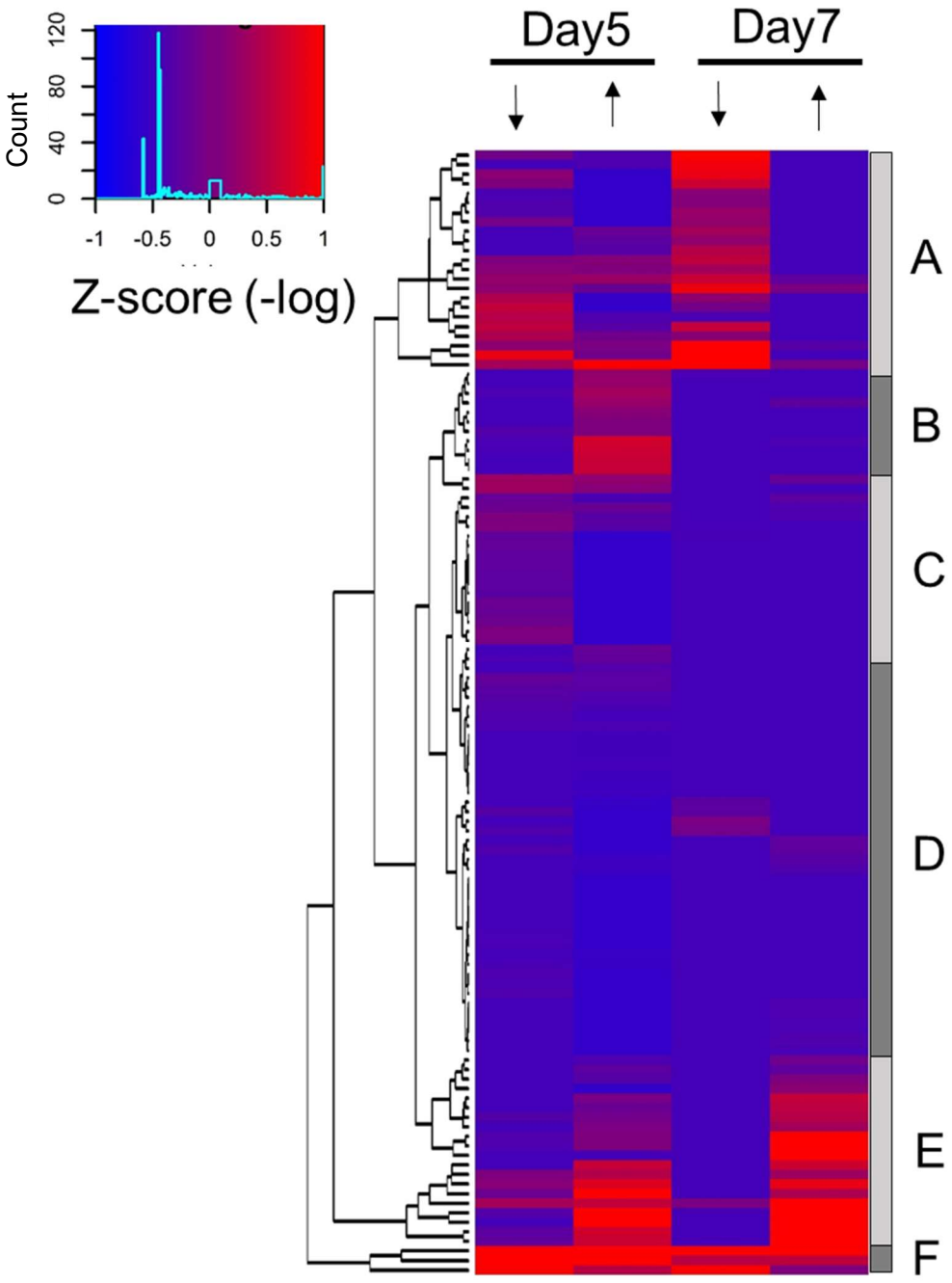


Figure 3.2. Impact of MESP1 on GO Developmental Processes. Map shows the z-score of the –log (P-value) of GO terms associated with upregulated and down regulated MESP1 target Genes between YFP + and YFP – cells from Days 5 and 7 EB differentiation. Clustering by “Complete” Method. Lists of GO terms in cluster groups can be found in Appendix C .

Table 2.1 Gene ontology analysis of Mesp1-repressive targets

Development GO terms enriched for MESP1 repressed targets			
GOID	Term	Count	Pvalue
GO:0007399	nervous system development*	111	4.66E-011
GO:0007423	sensory organ development	41	8.38E-007
GO:0007398	ectoderm development	24	3.30E-005
GO:0030902	hindbrain development	16	3.65E-005
GO:0048666	neuron development	40	4.50E-005
GO:0007420	brain development	39	7.30E-005
GO:0007417	central nervous system development	46	1.11E-004
GO:0001944	vasculature development	31	0.001919
GO:0001501	skeletal system development	34	0.002097
GO:0001568	blood vessel development	30	0.002611
GO:0043588	skin development	8	0.003498
GO:0021517	ventral spinal cord development	7	0.003738
GO:0031175	neuron projection development	27	0.004033
GO:0030900	forebrain development	22	0.005036
GO:0021549	cerebellum development	8	0.008574
GO:0021510	spinal cord development	9	0.011983
GO:0035295	tube development	29	0.013851
GO:0021675	nerve development	6	0.021285
GO:0021915	neural tube development	12	0.025688
GO:0031016	pancreas development	7	0.038179
GO:0007498	mesoderm development	9	0.046908
GO:0021681	cerebellar granular layer development	3	0.058446
GO:0001822	kidney development	13	0.061976
GO:0055123	digestive system development	6	0.079772

Count: number of genes associated with the GO term

Table 2.2. Repressed Nervous System Developmental Genes

Nervous System Development GO (GO:0007399)	
Day 5 Down Regulated (p = 4.66E-01)	Day 7 Down Regulated (p = 8.51E-006)
Shh (0.007) Mnx1 (0.03) Wnt3a (0.03) **Neurod1 (0.05)** Rax (0.05) Atp2b2 (0.06) Olig1 (0.06) Ihh (0.07) Ghrh (0.07) Lif (0.07) Olig2 (0.07) Rfx4 (0.07) **Foxa2 (0.09)** Chrd (0.11) Nlgn1 (0.11) Ntn1 (0.11) Slc1a2 (0.11) Fgf2 (0.13) Prrt1 (0.13) Six3 (0.13) Cobl (0.14) **Sox2 (0.15) Sox3 (0.15)** Ugt8a (0.15) Foxn4 (0.15) Erbb3 (0.15) Ccr4 (0.16) Foxp2 (0.16) Erbb4 (0.16) Elavl3 (0.17) Prdm13 (0.17) Agt (0.17) Bcan (0.17) En1 (0.17) Nfasc (0.17) Lhx3 (0.18) Pou4f3 (0.18) Rasgrf1 (0.18) Cntn2 (0.18) Jag2 (0.18) Mbp (0.18) Nodal (0.18) Phox2b (0.18) Wnt7a (0.18) Epha2 (0.19) Kndc1 (0.19) Nefm (0.19) Six1 (0.19) Gpr98 (0.21) Dtx1 (0.21) Foxd1 (0.21) Nefl (0.21) Dbx1 (0.22) Sepp1 (0.22) Grin2a (0.22) Myo6 (0.23) Zic1 (0.23) Ndrg2 (0.24) Lmx1b (0.25) Dpysl3 (0.25) Fabp7 (0.25) Ntrk2 (0.25) Clic5 (0.26) Alcam (0.27) Grin1 (0.27) Dmbx1 (0.28) Esr2 (0.28) Clu (0.28) Gnat1 (0.29) Itga3 (0.29) Bmpr1b (0.30) Pgap1 (0.30) Hoxc10 (0.32) Sema4a (0.33) Trp63 (0.33) Tomt (0.33) 4930506M07Rik (0.34) Tlx1 (0.34) Klk8 (0.34) Gsc (0.36) Myt1l (0.36) Hexb (0.37) Cabp4 (0.38) Slc1a3 (0.38) Btg2 (0.39) Ngfr (0.41) Nrn1 (0.41) Pax6 (0.41) Sema6a (0.41) Artn (0.42) Bai2 (0.42) Mecom (0.42) Hes5 (0.43) Phox2a (0.43) Atl1 (0.43) Pcdhb16 (0.43) Tubb3 (0.43) Dok5 (0.44) Pla2g10 (0.44) Egfr (0.45) Gbx2 (0.45) Jag1 (0.45) Cck (0.46) Egr2 (0.46) **Neurog3 (0.46)** Plxnb1 (0.46) Dab1 (0.47) Rorb (0.48) Tgif1 (0.49) Chl1 (0.49) Psen2 (0.49)	Mnx1 (0.04) Trp63 (0.06) Pla2g10 (0.08) Gabrr2 (0.09) Sphk1 (0.09) Atp2b2 (0.10) Lingo1 (0.10) Rgma (0.11) Esr2 (0.11) Sema4a (0.11) Wnt1 (0.11) Drd1a (0.12) Runx3 (0.13) Sox10 (0.14) Sox8 (0.14) Vsx2 (0.14) Dync2h1 (0.15) Gbx2 (0.15) Prrt1 (0.15) Lmx1b (0.17) Wnt7a (0.17) Ugt8a (0.18) Fabp7 (0.20) Gas1 (0.21) Insc (0.21) Shh (0.21) Egr3 (0.22) Nodal (0.22) Ntng2 (0.23) Bcan (0.23) Vwc2 (0.23) Gsc (0.24) Nkx6-1 (0.25) Agt (0.25) Gfi1 (0.26) Mdga1 (0.27) Tbr1 (0.27) Atoh1 (0.27) **Neurod1 (0.27)** Foxd1 (0.29) Ihh (0.30) Slitrk5 (0.30) Sema4g (0.31) Brsk2 (0.32) Olig2 (0.32) Pax7 (0.32) Pou3f2 (0.33) Chrd (0.34) Glrb (0.34) Sepp1 (0.35) Lhx2 (0.36) Erbb3 (0.36) Isl1 (0.37) Nkx2-2 (0.37) Slit1 (0.39) Six3 (0.40) **Neurog1 (0.41)** Pou3f3 (0.42) Rbm45 (0.42) Lhx5 (0.43) Phox2b (0.44) Barhl1 (0.45) Ntng1 (0.45) Thy1 (0.45) Elavl3 (0.46) Ngfr (0.46) Sema6c (0.46) Tcf7l2 (0.46) Fev (0.47) Chrnb2 (0.47) Nfasc (0.47) Dab1 (0.48) **Neurog2 (0.49)**

Fold Change (YFP + / YFP -)

3.2.3. MESP1 suppresses key neuro-developmental genes

While in-silico analysis that compared MESP1-targets to RNA-seq give us cogent correlative evidence, I next wanted to investigate whether forced MESP1 expression would also diminish the expression of neuroectoderm targets. I introduced a doxycycline (DOX)-inducible tet-on expression vector of GFP-tagged MESP1 into mouse embryonic stem cell AB2.2. The mES cells cultured in monolayers were induced by DOX, FAC sorted for YFP, and collected for each data time point, to guarantee that non-induced cells were not excluded from the Dox-induced population (Figure3.3). *Mesp1* mRNA levels increased 400 fold within the first six hours of Dox treatment compared to the zero time point (Figure 3.3C). The expression levels of *Sox2, Neurod1, Neurog1, Neurog2,* and *Neurog3*, showed significantly reduced mRNA levels in the *Mesp1*-induced population over non-induced cells (Figure 3.3D-H). This is in contradistinction to the rapid induction of factors, such as *Gata6, Gata4,* and *Pitx2*. Thus, MESP1 plays a repressive role on the transcription on neuro-ectodermal-dependent transcription factors, likely in a direct-dependent manner.

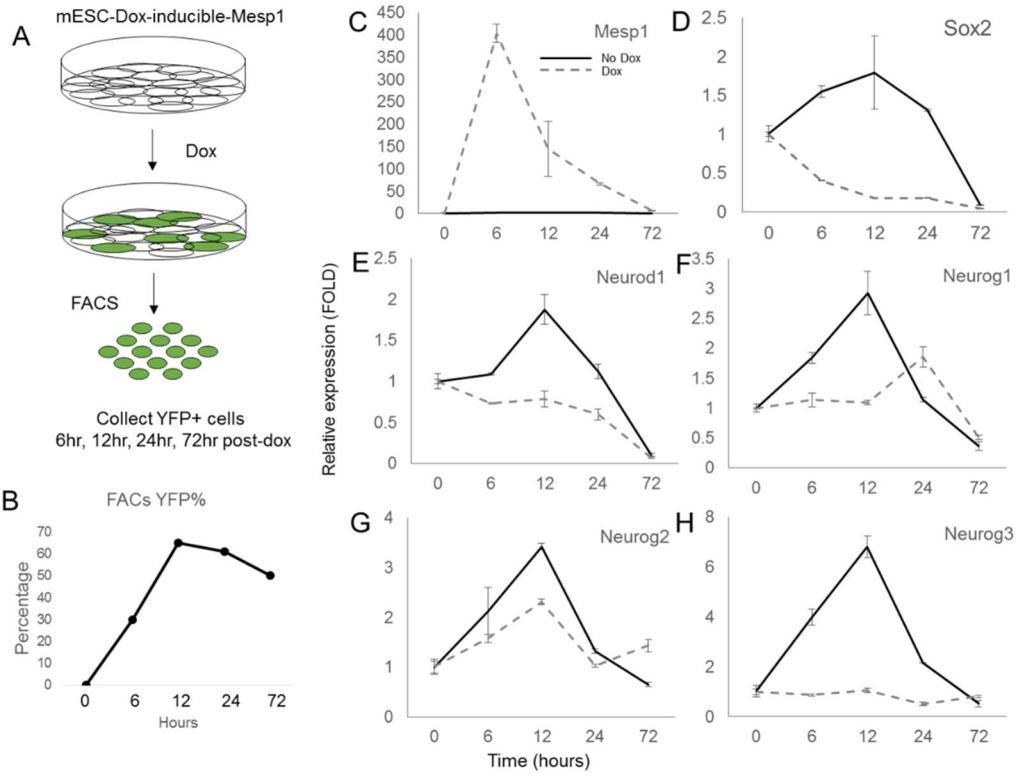


Figure 3.3. MESP1 Represses Neuro Specific Gene Targets. A) Diagram of Experimental design. Dox inducible MESP1 mouse embryonic stem cells will be treated with a onetime dose of 1 ug/ml of Dox and RNA were extracted from FAC-sorted cells at time points 6, 12, 24, and 72 hours. B) YFP percentage collected post Dox induction for each time point. C-D) Real-time quantification of *Mesp1, Sox2, Neurod1, Neurog1, Neurog2,* and *Neurog3*. Data represents relative expression compared to undifferentiated mES cells. Error-bars represents standard deviation of technical replicates (n=3), repeated (n=3) data not shown. RT-qPCR show that MESP1 represses the expression of target genes with in the first 12 hours of expression.

3.2.4. MESP1 and H3K27me3 binding sites enrichments overlap during early developmental stages

Regulation of gene transcription can occur at multiple levels from the formation of heterochromatin for a sustained suppression to a transient regulation utilizing H3K27me3 through the PRC2 complex (Margueron and Reinberg 2011, Blackledge, Rose et al. 2015, Khan, Lee et al. 2015). Next, I wanted to determine if MESP1 regulated key neuro-ectodermal transcription factors were affected at the level of epigenetic marks by testing the deposit of H3K27me3. A genome-wide study of different histone modification into cardiac lineage using mouse embryonic stem cells has been previously characterized and deposited in public data bases (Wamstad, Alexander et al. 2012). First, I wanted to determine if a trend existed during the transition from mES cells to mesendodermal cells. I quantified enrichment of H3K27me3 tags over MESP1 peaks by tag density utilizing HOMER annotatePeak.pl program with given MESP1 peak size. I found that among the 217,565 MESP1 peaks, 135,489 peaks had an increase in H3K27me3 enrichment, and 79,087 peaks had a decrease in enrichment (Figure 3.4A), which indicated a potential recruitment role for MESP1 for the PRC2 complex.

I observed an overall increase in H3K27me3 modification overlapping Mesp1-binding sites associated with neuro-specific GO terms and a general decrease over Mesp1-binding sites associated with Meso-cardiac specific GO terms (Figure 3.4B). I observed that the chromatin landscape for MESP1 binding overlapped well with increased H3K27me3 enrichment on key neuronal-specific

developmental regulated genes, *Neurod1, Neurog1, Neurog2,* and *Neurog3* (Figure 3.4C-F).

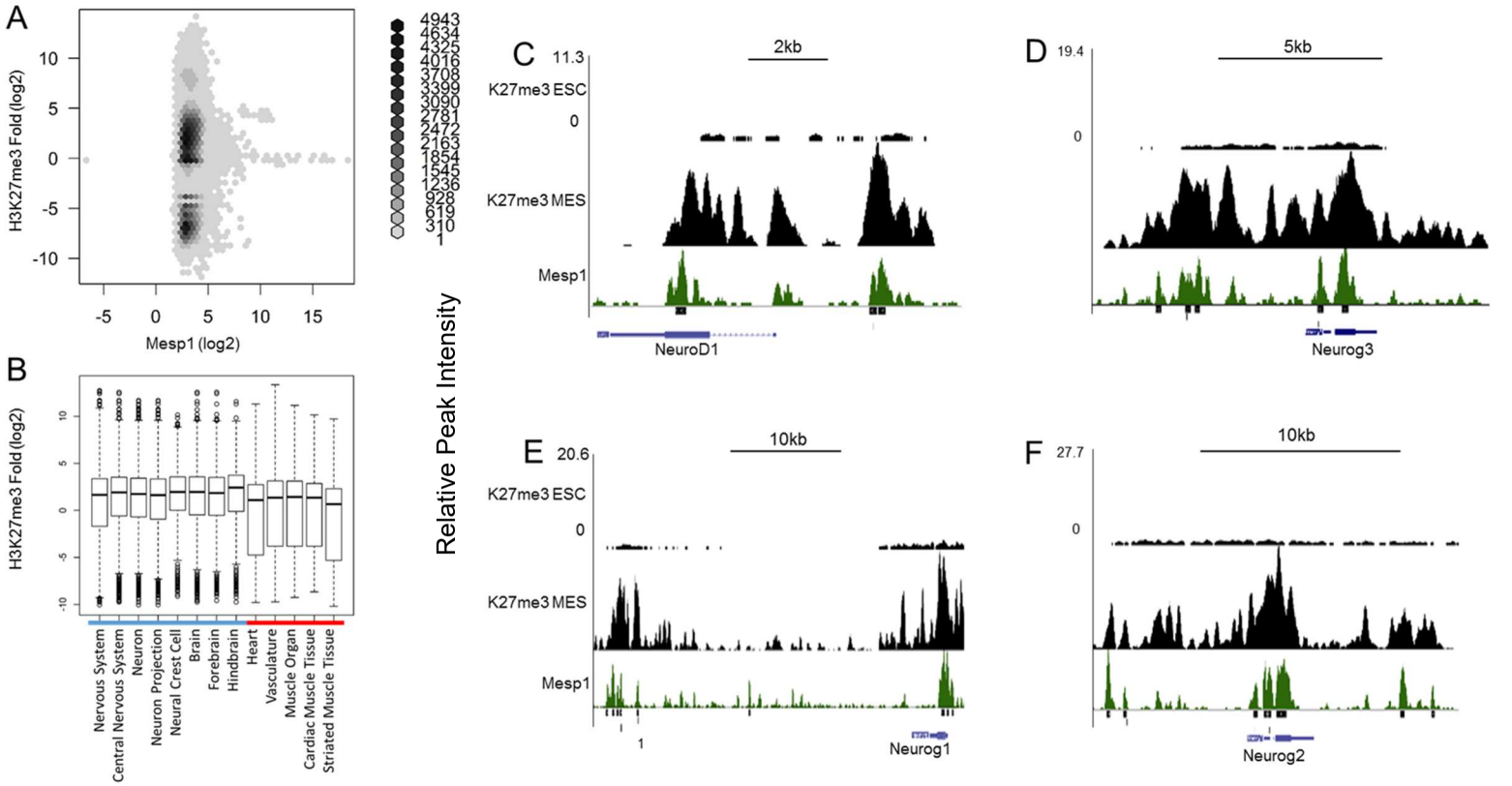


Figure 3.4. MESP1 and H3K27me3 comparison. A) Correlation density plot between the MESP1 peaks and H3K27me3 fold change (MES/ESC). MESP1 peaks experience both an increase and decrease in H3K27me3 deposit. B) Fold change of H3K27me3 over MESP1 peaks associated with specific GO terms. MESP1 peaks associated with Neuro GO terms have an increase in H3K27me3 deposit, while the Cardiac muscle GO terms experience a degree in H3K27me3 deposit. C-F) UCSC bedgraph of H3K27me3 ChIP-seq data overlapping MESP1 ChIP-seq data. H3K27me3 modification overlaps MESP1 binding sites that are associated with Neurod1, Neurog3, Neurog1, and Neurog2 enhances and promoter regions.

3.2.5. MESP1 binds and recruits the PRC2 complexes

Next, I asked whether MESP1 directly influences the deposit of H3K27me3 on targeted sites through the polycomb group complex. I searched within the promoter and enhancer region of the highly affected neuronal gene, neurog3 and identified three eboxes that lie within MESP1 peaks (Figure 3.5A). To test my hypothesis I used our dox-inducible MESP1 expressing YFP marked mouse embryonic stem cells in which induced MESP1 bound and recruited to the PRC2 complex to the binding site 1, which increased the levels of H3K27me3 modification. Ebox site 2 showed a slight increase in MESP1 binding of the polycomb PC2 protein, EZH2, while H3K27me3 deposit barely showed any change on the third Ebox.

I further analyzed the three different binding sites to search for any intrinsic difference that would suggest a bias in ebox one over the two other eboxes. I found that the first ebox was a dissimilar variant compared to the other two eboxes (Figure 3.5B-C) (CACCTG and CAGCTG). Both variants are Class I eboxes associated with Class A bHLH transcription factors, in which MESP1 is also a member (Ledent and Vervoort 2001, Jones 2004). Could MESP1 have preferential binding between the two variants eboxes where on is preferential for repressive activity? Note, ebox binding selectivity on DNA is within its bHLH domain, determined primarily by its basic domain.

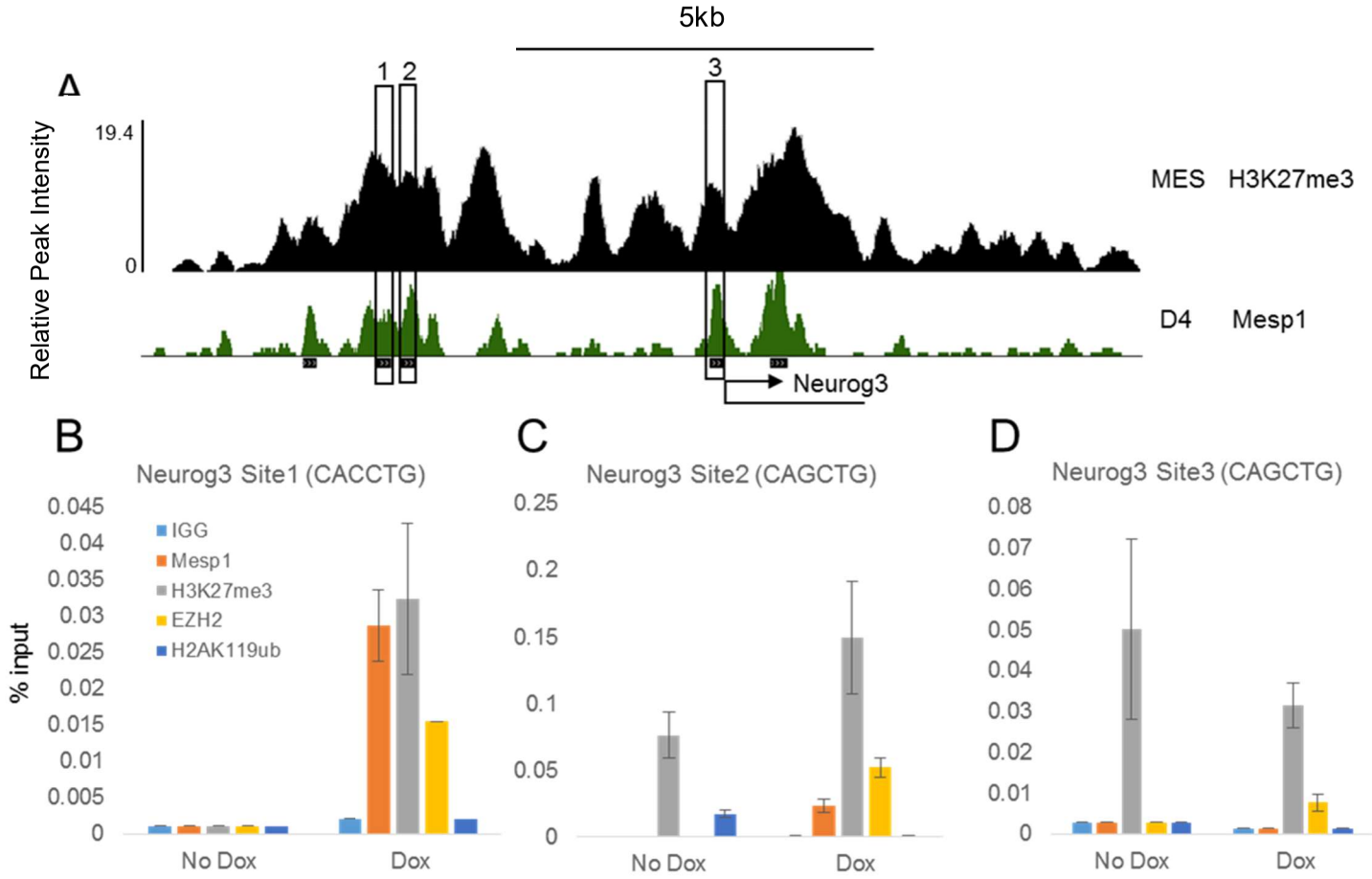


Figure 3.5. Recruitment of PRC2 complex on neurog3 enhancers by MESP1. A) UCSC bedgraph image of Neurog3 enhancers and promoter regions showing MESP1 D4 ChIP-seq Peaks over lapping with H3K27me3 MES ChIP-seq Peaks. Box regions indicate locations of ebox motif where PCR primers was designed for ChIP-qPCR analysis. B-D). ChIP-qPCR of IGG, MESP1, H3K27me3, EZH2, and H2AK119ub over Neurog3 sites 1, 2, and 3 with the induction of MESP1 in mouse embryonic stem cells for 12 hours. Each site contains one ebox with different variants. No Dox represents negative control. Error bar = Standard deviation of three technical replicates, repeated (n=3), data not shown.

Using in-silico analysis of crystal structures from different bHLH ebox DNA complexes, a conserved E/(RK)N(R/N) motif across the bHLH family has been identified to be responsible for selectivity (Atchley and Fitch 1997, Ledent and Vervoort 2001, Zheng, Zheng et al. 2009, Masi, Grove et al. 2011). To determine if MESP1 preferentially binds to one variant, I mutated lysine residue 86 within the binding motif (Appendix C Figure S2), which will potentially disrupt MESP1's ability to bind to the phosphate backbone of the ebox, resulting in changes in DNA binding specificity (Masi, Grove et al. 2011). I also generated neurog3 sites 1 (CACCTG) and 2 (CAGCTG) DNA probes to determine if mutant MESP1 would have differential binding. When each probe was incubated with a cell lysate containing MESP1 and mutants from 293T cells, wild type MESP1 showed preferential binding to neurog3 site 2, but when mutated there is a loss of affinity to site 2 and an increased affinity to site 1 (Figure 3.6). The ability for MESP1 to recognize and bind to neurog3 site 1 may require a heterodimer complex with a different bHLH.

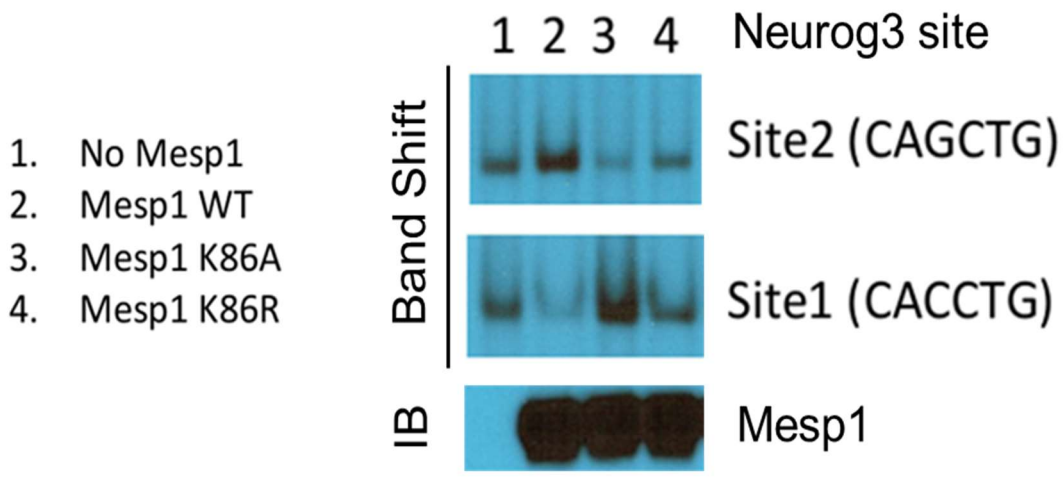


Figure 3.6. Lysine 86 determines binding specificity. Northern Blot for MESP1 incubate with P^{32} labelled DNA probes of the two Neurog3 binding sites show differential binding between the WT, K86A, and K86R MESP1 protein. Immunoblot of Lysate used show even expression of MESP1 protein.

3.2.6. MESP1 utilizes variant eboxes based on transcriptional function

To elucidate globally the potential for MESP1 to bind different motifs, with differential functions, I searched *in-silico* within the MESP1 enriched region for ebox motifs (CANNTG) using HOMER motif analysis. The ebox motif has eight potential variants due to the flexibility of the center two bases, NN and the possibility that each variant may pertain to a certain function in respect to the different bHLH transcription factor (Soleimani, Yin et al. 2012). I searched within all of MESP1 peaks for all possible ebox variants. This resulted in two dominant variants CAGCTG made up 20.46% (p-value $1.00x10^{-301}$) of the enriched regions, and CACCTG made up 25.67% (p-value $1.00x10^{-31}$) similar to previously published data on MESP1.

To determine the possibility that MESP1 has preferential binding and differential function by ebox variants, I filtered MESP1 peaks first by locating and counting the different ebox variants, then I separating the count by set of peaks per gene. After the count is separated I grouped the counts by the fold change of the target gene between YFP+/- cells. I ran the same filtering flow-through with peaks with only H3K27me3 modification. I found that among the total MESP1 enriched regions there is an average of 3.46 (30.39% average per gene) CACCTG ebox variant per gene compared to 2.76 (23.62% average per gene) CAGCTG ebox variant per gene (t-test paired, $5.75x10^{-52}$), indicating that MESP1 may favor CACCTG over CAGCTG globally (Figure 3.7).

Next I determined if a particular variant is a more or less abundant when the target gene is active or repressed, the ebox peaks were grouped accordingly with the RNA expression levels of annotated TSS. For enriched regions that are adjacent to a gene that experienced an increase in RNA expression, I also observed an increase in CAGCTG ebox variant per gene to 3.05 and no changes in the average CACCTG variant, suggestive of a bias towards CAGCTG over the CACCTG variant. When analyzing peaks associated with genes that are reduced by 0.5 fold I observed an average per peak of 3.79 (32.01% average per gene) of CACCTG and an average per peak appears of 2.88 (26.82% average per gene) of CAGCTG (t-test paired, $1.02x10^{-54}$). The increase in CACCTG over CAGCTG for genes that where repressed would suggest a repressive bias for CACCTG variant (Figure 3.7).

Futhermore, I asked whether the bias for CACCTG over CAGCTG in respect to repression is maintained when MESP1 binding overlapes with H3K27me3. I found that all of the MESP1 enriched regions that have overlap with H3K27me3 during the mesoderm stage have an average per gene count of 1.56 for CACCTG and an average per gene count 1.22 CAGCTG. Thus, still maintaining a bias towards CACCTG in total MESP1-bound targeted regions (t-test paired, $4.81x10^{-82}$) (Figure 3.6). Again to determine if the variant distribution is affected by gene expression, enriched regions are separated by the fold change. I found that MESP1-enriched regions that were associated with genes that are upregulated (Fold >2.0) have no significant change in ebox variant distribution seen when comparing all enriched regions (CACCTG = 1.89 average per gene, CAGCTG =

1.72, t-test paired = .10). For genes that are repressed (Fold < 0.5), I observed an increase in both variants compared to the total regions, but a greater difference between variant (CACCTG = 1.98 average per gene, CAGCTG = 1.41 average per gene, and t-test paired = 8.06×10^{-41}) (Figure 3.7).

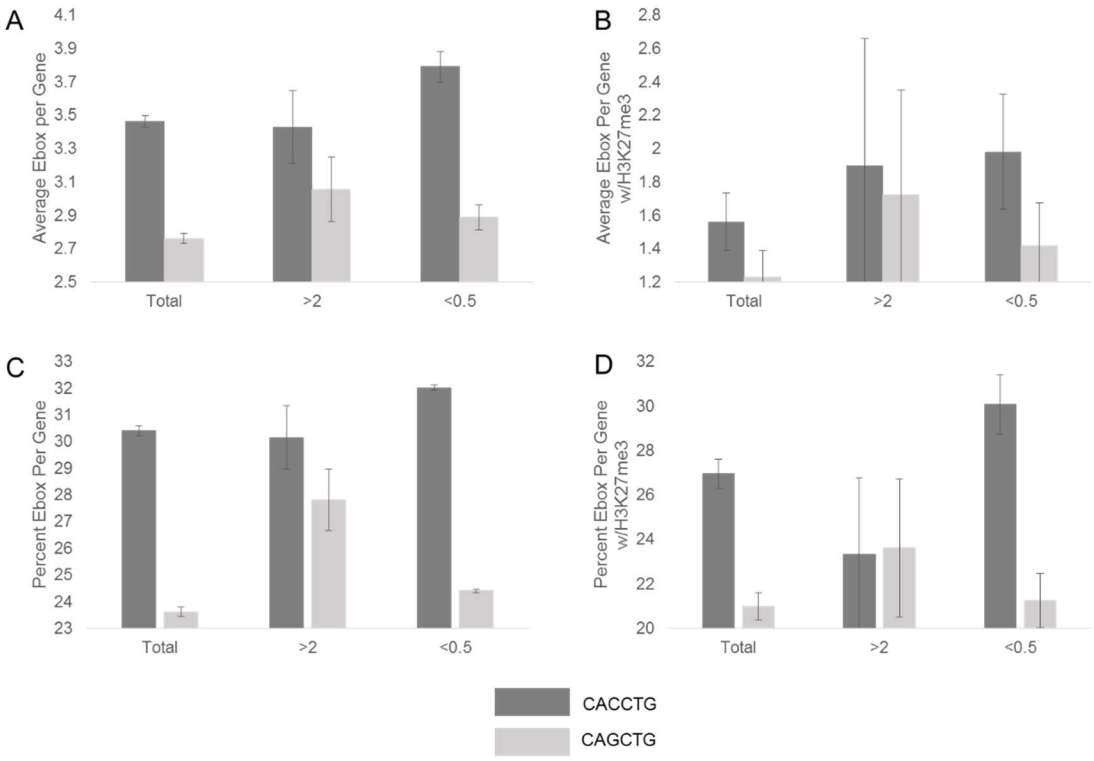


Figure 3.7. Ebox variant count and percentage in MESP1-binding sites. A,B) The average count of CACCTG and CAGCTG ebox variant per gene in total Mesp1-binding sites. A) Total Mesp1-binding sites. B) Overlapping with H3K27me3. C,D) Percentage of ebox variant relative to the number of Peaks associated to target gene. C) Total Mesp1-binding sites. (D) Overlapping with H3K27me3 modification. Error bars = standard error mean of ebox variant per set of Mesp1-binding sites of one biological replicate. Analysis repeated with separate independent biological replicate, (n=2) data not shown.

3.3. Discussion

The formation of the mesoderm germ layer is a dynamic multifactorial process that occurs during the expression of MESP1 and other mesendoderm markers, T, Eomes, and Goosecoid at 6.5 dpc in the mouse. The process itself required a delicate balance of multiple lineage markers at the surface and nuclear level that controls the transition of the epiblast into the three specific germ layers (endoderm, mesoderm, and ectoderm).

In order for mesoderm to properly form, it is important that Eomes (Costello, Pimeisl et al. 2011), and T (David, Jarsch et al. 2011) are expressed at the proper time, in turn activating MESP1. The activation of MESP1 begins the EMT transition process that allows the migration of the nascent mesoderm from the primitive streak. Controlled expression of MESP1 is critical for the viability of the embryo, if failure of activation occurs, lethality is seen by 8.5-9 dpc due to inhibition of EMT transition and malformation of the heart resulting in cardia bifida (Saga, Miyagawa-Tomita et al. 1999). Inversely, the transcript of MESP1 is present for a very short period of time, and the inhibition to control the expression will also result in an embryonic-lethality failure of proper differentiation of the posterior mesoderm layer. The lethality of MESP1 KO studies has established the importance of completely understanding the mechanism and role it plays during mesendoderm formation and maturation.

The repression of the epiblast/ectoderm signature during gastrulation is important for the proper formation of both the endoderm and mesoderm layers and

our evidence suggests that MESP1 may play a role, in addition, the importance of properly inhibiting the epiblast/ectoderm differentiation factors hasn't been well-studied. During embryogenesis, MESP1 progenitor cells contribute to the cranial-cardiac mesoderm and the paraxial mesoderm tissues, which are directly adjacent to the notochord and the neuro-fold.

Our previous finding analyzing the transcriptome of MESP1 progenitor cells, revealed a potential for MESP1 to act as a transcriptional repressor on non-mesoderm lineages, in addition, recent findings show that MESP1 can indirectly regulate the neuro-ectoderm program through the expression of miRNA's that block the expression of Celf1 (Shen, Soibam et al. 2016). A re-interpretation of our previous findings has shed light on MESP1's direct potential to regulate the neuro-specific ectoderm program. This may have been overlooked in previous studies due to its implication in the formation of the heart, and that lack of contribution of MESP1-fated cells into neuro-ectoderm lineages.

The potential dual function of MESP1 and the motif analysis of two predominate ebox variants lead us to believe that its function is dependent on the variant. Our filter analysis based on MESP1-enriched regions relative to RNA expression levels between MESP1-expressing and-non expressing cells of adjacent genes suggest that MESP1 utilizes the CACCTG variant over CAGCTG to repress gene targets. Further filtering utilizing histone modification H3K27me3 for repression and H3K27ac for activation resulted in an inverse relationship relative to its chromatin state. MESP1-binding to neurog3 at CACCTG in site 3

recruited EZH2 to that location and deposited methyl groups to Histone 3 lysine 27. Co-immunoprecipitation of MESP1 and the PRC2 complex will determine if MESP1 directly interacts with a component of the complex. However, no evidence was observed. We also chipped for PRC1 complex ring1 and H2A119ub to determine if the binding region is signaled to convert into heterochromatin during the expression of MESP1. However, we observed no deposit which would indicate that MESP1 only can only transiently repress transcription with the deposit of H3K27me3. The conversion of euchromatin to heterochromatin may require downstream factors that replace MESP1 after its expression is reduced.

The failure to repress MESP1's neuro-ectoderm targets through the PRC2 complex may cause malformation in the development in the neuro-fold, neuro-tube, and the axial, paraxial mesoderm, due to its proximal location during embryogenesis. Further studies on the effects of MESP1 on the PRC2 complex will need to be investigated. In addition, due to the early lethality of MESP1 KO, and failed migration to form the cardiac and crania-mesoderm, improper conversion of crania-mesoderm into notochord or neuro-fold in the anterior region cannot be studied.

4. Chapter 4 Overview

It is important to understand the reductive nature of MESP1 and how other transcription factors may interact and contribute regulatory activity in directing somatic and stem cell reprogramming. It is also important to understand its impact on biological or developmental processes but the mechanism of transcription factors like MESP1 on a global level with whole transcript and DNA-binding sequence analysis. It is critical not to rely heavily on the gain of function analysis with forced expression that can result in many false positives. A holistic approach is better to determine transcription factors targets endogenously and their effects during differentiation through lineage tracing during differentiation. The focus of this study was to determine MESP1's endogenous function, its effects on the cardiac program and how it manipulates the chromosome landscape at target sites.

At the beginning of this study, MESP1 was determined to be at the top of the hierarchy tree in terms of initiating the cardiac program, with subsequent activation of *Mef2c, Nkx2.5, Myod, Hand2, Gata4, and Gata6*. (Figure 2.2, 2.4, 2.4, 3.2, and Table 2.1) (Saga, Miyagawa-Tomita et al. 1999, Kitajima, Takagi et al. 2000, Saga, Kitajima et al. 2000, Lindsley, Gill et al. 2007, Bondue, Lapouge et al. 2008). However, the comparison between the gain of function investigations of MESP1 resulted in conflicting conclusions, which leads us to determine if it directly or indirectly activates the cardiac gene program. (Lindsley, Gill et al. 2007, Bondue, Lapouge et al. 2008). The hurdle to overcome was to distinguish whether MESP1

directly or indirectly regulates target cardiac genes and its effects on the whole developmental program. It is accepted that MESP1 can cell-autonomously activate the cardiac program in undifferentiated and differentiating mES cells by activating the core cardiac program. However, the studies done previously on MESP1's regulatory role is subject to potential artifacts from transient over expression which is evident compared to these findings.

This study shows that MESP1-marked cells eventually activate the core cardiac program but after 48 h after the activation and degradation of MESP1 (Figure 2.4). This is evident from *in vivo* embryo and mES cell EB differentiation studies conclude that the expression of mRNA of MESP1 is quickly down-regulated after expression on day four of EBs and in the lateral migration of mesoderm cells during embryogenesis (Figure 2.1).

This gap in time between MESP1 expression and the activation of the core cardiac program (Figure 2.4C, D, 2.5A) showed that the activation is, in fact, indirect. And the potential process in which the signal is relayed through epigenetics is evident, but, however, not conclusive. (Figure 2.6 and Figure 3.4). The limitation of this study is that it is correlative between the stages of mES cell differentiation and does not determine causality. It would be necessary to utilize both gain and loss of function to properly determine causality in respect to MESP1 and changes in epigenetic through chromatin remodeling.

In addition, it is evident that MESP1 is more than just an activator of the core cardiac program and initiator of EMT, which was the focus of MESP1's

function in the past decade, but a repressor of non-mesoderm programs, and a chromosome modifier that it has a lasting effect that transcends its presence. The function of MESP1 has been evident in previous literature, showing that it can affect the expression of key endoderm factors like Foxa2 and Sox17 (Bondue, Lapouge et al. 2008). It has been suggested that MESP1 may repress endoderm formation during gastrulation, however, our evidence suggests that this isn't completely true (Figure 2.4, Figure 2.5, and Appendix B Figure S4). Many early endoderm developmental programs are activated within MESP1 marked cells and co-staining show the expression of T, Foxa2, and Sox17 with YFP.

The repressive function of MESP1 is further expanded in this study and the mechanism through the PRC2 complex that is used. We found that during differentiation, MESP1-marked cells show a decrease in neuro-ectoderm factors which is expected, however, the analysis of all of MESP1's potential targets show an overwhelming association with neuro-ectoderm developmental gene targets rather than those representing mesoderm and endoderm-expressed genes. This finding is supported by Lindsley et al. (Lindsley, Gill et al. 2007), utilizing DKK1 and a WNT inhibition which increased the expression of neuronal factors in mES cells repressed by MESP1.

Furthermore, dual roles of MESP1 lead us to investigate potential DNA binding preferences. In this study, I revealed new supporting evidence that individual ebox DNA motifs, may have Mesp1 binding favorites, biased with one variant that coincides with function. The notion that MESP1 may utilize a specific

ebox variant opens the door to new research in regards to understanding the complex nature of all bHLH factors.

It is evident from this study that MESP1 changes binding affinity depending on the molecular makeup of the cell. In 293T cells the expression of only Mesp1 allowed for the binding to CAGCTG as a homodimer, but when K86 of MESP1 was mutated to alanine and arginine, the mutation of MESP1 altered affinity to CACCTG (Figure 3.7). However in mouse ES cells, AB2.2, MESP1 can bind to neurog3 site 1 that contains the CACCTG ebox variant and recruit the PRC2 complex and deposits methylation onto K27 of histone 3 (Figure 3.5). This suggests that in order to achieve MESP1 binding and recruitmenting of the PRC2 complex onto site 1 of the neurog3 site, MESP1 likely heterodimerizes with a different bHLH factor that is present in mES cells but not in 293T cells. In addition, bHLH transcription factors act in tandem with other bHLH factors and primary function as a heterodimer. This heterodimer cofactor potential may disrupt Lysine 86 binding to the phosphate backbone of the DNA helix allowing for increased affinity for ebox binding. (Skinner, Rawls et al. 2010, Masi, Grove et al. 2011).

A phylogenetic analysis of mouse bHLH factors has divided all known factors into groups based on total sequence similarities. MESP1 belongs to a family identified as Clades C. This family contains MESP2, FIGLA, MYF6, MYOG, MYOD1, MYF5, MXD3, MAD, MXL1, MXD4, MAD4, and TRAFP4. Clade C factors have the propensity to co-factor with Clade D factors, which include, TCFL4, MNT, SREBF1, SREBF2, and MAX. In addition, the lysine 86 residue found in MESP1

that determine binding specification can also be found in other Clade C MAD factors, which are known to co-factor with Clade D MAX factors, to repress target genes (Masi, Grove et al. 2011). This leads to the possibility that MESP1, when co-factored with MAX factors, will bind and repress target genes by recruited the PRC2 complex. Further investigation is required to show that this is, in fact, the case when it comes to MESP1 repression.

With the improvement of ChIP and sequencing technology, is has become increasingly possible to analyze the DNA binding sites of various transcription factors like MESP1 from a relatively small sample size. To understand a transcription factors unbiased or natural binding sites, analysis of endogenous expressing protein is critical to identify its function. In addition, an added benefit in studying a transcription factors in its endogenous level and proper stage in development is that the necessary co-factors that is required for a function that may not be present in model systems like 293T. This, however, does not discredit the majority of findings using forced expression but should be approach with much skepticism.

I conclude from my study that; 1) MESP1 indirectly activates the core cardiac program, *Nkx2-5, Mef2c, Hand2,* and *Myod1* potentially through histone modification by H3K27acetylation and relays the activating signal through downstream factors, Hand2, Twist, and Gata factors, 2) MESP1 directly represses the non-mesodermal programs like neuro-ectoderm through the PRC2 complex

resulting in the deposit of H3K27me3. 3) MESP1 directs function utilizing different ebox variants through potential co-factors.

5. Appendix A Materials and Methods

Immunostaining: For immunostaining of differentiated cardiomyocytes, chamber slide cultures were fixed in 4% paraformaldehyde for 10 minutes, then permeabilized in 0.1% Triton X-100 in phosphate buffer saline (PBS) for 30 minutes at room temperature. Next, the slides were incubated in blocking buffer (10% normal goat serum, 0.1% Triton X-100 in PBS) for 30 minutes at 37^0C, then in 1:100 anti-a-actinin antibody diluted in blocking buffer at 4^0C overnight. The slides were next incubated in 1:500 fluorescein isothiocyanate (FITC)-conjugated anti-mouse antibodies for 90 minutes at room temperature. Finally, the slides were mounted in DAPI-containing mounting media and documented under a Nikon fluorescent microscope.

For co-staining of T, Foxa2, or Sox17 with YFP, day five EBs were used, and a similar staining procedure as above was followed. The primary antibodies used were T (Santa Cruz Biotech, H-210), Foxa2 (Santa Cruz Biotech, H-150), or Sox17 (Santa Cruz Biotech, H-130) together with biotin-conjugated anti-GFP (Novus). The secondary antibodies used were FITC-conjugated goat anti-rabbit (Molecular Probes) and allophycocyanin (APC)-conjugated streptavidin (Molecular Probes).

Karyotype: Karyotyping was performed by The Texas Children's Cancer Center Cytogenetic Research Lab using the Giemsa staining method on metaphase chromosomes.

Western Blot: Samples were lysed in radio immunoprecipitation assay (RIPA) buffer supplemented with protease inhibitors (P8340, Sigma Aldrich) and the protein concentrations were determined by BCA Protein assay kit (PI-23225, Thermo Scientific). The samples were denatured by boiling for five minutes with NuPAGE LDS Sample Buffer (NP0007, Life Technologies), and subjected to electrophoresis with Novex NuPAGE system. Then the proteins were transferred to polyvinylidene difluoride (PVDF) membranes. The PVDF membranes were blocked with 5% TBST-milk before treatment with the primary antibody and followed by horseradish peroxidase (HRP) conjugated secondary antibody incubations. Enhanced chemiluminescence (ECL, NEL104001EA, Perkin Elmer) reagent was used to visualize the targeted protein. The antibodies used were as follow: MESP1 (rabbit polyclonal, Bethyl Lab) and GAPDH-HRP (sc-20357-HRP, Santa Cruz Biotechnology).

RNA Isolation and qRT-PCR: In the UH3 characterization studies (Appendix B Figure S1), unsorted EB RNA was isolated using Trizol (Life Technologies). RNA was extracted using a Direct-zol RNA MiniPrep kit from Zymo Research. Following RNA isolation, RNA was quantified using a Nanodrop and cDNA was synthesized using qScript cDNA Supermix from Quanta Bioscience. QPCR was performed using SYBR Select Mast Mix following the Manufacturer's guidelines. Primers were from the Primerbank database (Wang, Spandidos et al. 2012) unless specified otherwise.

The primer sequences with their respective Primerbank ID were:

T-forward 5’ GCTTCAAGGAGCTAACTAACGAG 3’

T-reverse 5’ CCAGCAAGAAAGAGTACATGGC 3’, 6678203a1;

MESP1-forward 5’ GCTCGGTCCCCGTTTAAGC 3’

MESP1-reverse 5’ ACGATGGGTCCCACGATTCT 3’, 254588103b1;

Gata4-forward 5’ CACGCTGTGGCGTCGTAAT3’

Gata4-reverse 5’ CTGGTTTGAATCCCCTCCTTC 3’, 110681730b3;

Tbx5-forward 5’ TGGCTGAAGTTCCACGAAGTG 3’

Tbx5-reverse 5’ TTTGGGATTAAGGCCAGTCAC 3’, 229577242b2;

Mef2c-forward 5’ ATGCCATCAGTGAATCAAAGGAT 3’

Mef2c-reverse 5’ GTGGTACGGTCTCCCAACT 3’, 13384624a1;

Nkx2.5-forward 5’ GACAAAGCCGAGACGGATGG 3’

Nkx2.5-reverse 5’CTGTCGCTTGCACTTGTAGC 3’, 6679068a1;

Sirpa-forward 5’ CACGGGGACAGAAGTGAAGG 3’

Sirpa-reverse 5’TGCAGTTGAGAATGGTCGAATC 3’, 6671640a1;

Cxcr4-forward 5’ GAAGTGGGTTCTGGAGACTAT 3’

Cxcr4-reverse 5’ TTGCCGACTATGCCAGTCAAG 3’, 6756460a1;

Flk1-forward 5’ TTTGGCAAATACAACCCTTCAGA 3’

Flk1-reverse 5’ GCAGAAGATACTGTCACCACC 3’, 27777648a1;

Pdgfra-forward 5’ GGAGACTCAAGTAACCTTGCAC 3’

Pdgfra-reverse 5’ TCAGTTCTGACGTTGCTTTCAA 3’, 134032015c3;

Alcam-forward 5’ CTCGTTGCTGGTGTCGTCTA 3’

Alcam-reverse 5’ AATCCGCTCCTCTCTTAGGC 3’

Gapdh-forward 5’ AGGTCGGTGTGAACGGATTTG 3’

Gapdh-reverse 5’ TGTAGACCATGTAGTTGAGGTCA 3’, 6679937a1.

Primers Designed by Blast-primer (NCBI)

Sox2 forward 5’ GCGGAGTGGAAACTTTTGTCC 3’

Sox2 reverse 5’ CGGGAAGCGTGTACTTATCCTT 3’

Neurog3 forward 5’ CCAAGAGCGAGTTGGCACT 3’

Neurog3 reverse 5’ CGGGCCATAGAAGCTGTGG 3’

Neurod1 forward 5’ ATGACCAAATCATACAGCGAGAG 3’

Neurod1 reverse 5’ TCTGCCTCGTGTTCCTCGT 3’

Neurog1 forward 5’ CCAGCGACACTGAGTCCTG 3’

Neurog1 reverse 5’ CGGGCCATAGGTGAAGTCTT 3’

Neurog2 forward 5’ AACTCCACGTCCCCATACAG 3’

Neurog2 reverse 5’ GAGGCGCATAACGATGCTTCT 3’

MESP1 ChIP-qPCR primers Designed by Blast-primers (NCBI)

Neurog3 Site1 forward 5’ GGAGCACCCGAGGTCTTTAT 3’

Neurog3 Site1 reverse 5’ GAAGAGGCGTGGGTTCAACA 3’

Neurog3 Site2 forward 5’ GTGGGGACAGATAAGGCGAG 3’

Neurog3 Site2 reverse 5’ TCATTCAAGAGCGCACAAGC 3’

Neurog3 Site3 forward 5’ CTGGCACGCTTTATCTGCTC 3’

Neurog3 Site3 reverse 5’ CCTGCCCTTTGTCCGGAAT 3’

The RNA of AB2.2 was isolated by RNeasy Mini kit (74104, Qiagen) and quantified using Nanodrop. The cDNA was synthesized using the M-MLV reverse transcription kit (28025-013, Life Technologies). Quantitative real-time PCR was carried out using Brilliant II SYBR Green QPCR Master Mix (600828, Agilent

Technologies). The primers used for *MESP1* and *Gapdh* were the same primers as indicated. In the remaining RNA-Seq and ChIP-Seq studies, RNA was isolated from undifferentiated, and FACS sorted UH3 cells at day 5-8 post- differentiation. Both the positive and negative populations were used for transcriptome analysis. RNA was extracted from 250,000 cells for each group using the RNAeasy kit from Qiagen following the manufacturer's recommended procedure. Following isolation, the RNA was quantified on a Nanodrop.

RNA-Seq Library Preparation: A total of 100 ng of the isolated RNA was used for generating the library for next-generation sequencing using Illumina's HiSeq platform at MD Anderson DNA core facility. Libraries were prepared using the ScriptSeq Complete Gold kit from Epicentre (Cat. No. SCL24EP) following the manufacturer's recommended procedure for low input starting material. Final libraries were quantified using the pico-green quantification kit from Promega using Promega's Quantifluor fluorometer.

RNA-Seq Data Processing: RNA-Seq reads were aligned to the mouse genome using tophat with parameters: -p 8 --read-mismatches 2 --b2-L 20 -g 5. Reads that mapped to multiple locations of the genome were discarded before downstream analysis. An annotated list of known transcripts (mm9 UCSC refseq genes) was used for a final set of the transcriptome. The abundance and number of raw fragments aligned to each gene were computed using cufflinks (Trapnell, Williams et al. 2010). The abundance of each gene was expressed as FPKM (Fragments Per Kilobase of transcript per Million mapped reads).

To obtain differentially expressed genes between two conditions, we used DESeq (Anders, Anders et al. 2010). The raw counts mapped to a transcript were obtained using htseq-count and were used as inputs to the DESeq tool. To remove genes with low expression profiles, we included only the ones with a total aligned fragment count across all three biological replicate samples greater than five. We used P-value of less than 0.05 as the criteria to determine differentially expressed genes between two conditions.

To perform hierarchical clustering of genes associated with different germ layers, we obtained genes which were annotated under direct GO terms – mesoderm development, endoderm development, and heart development. The robust z-score was computed for each gene using its FPKM values across ES, days 5, 6, 7, and 8 of YFP+ cells. We used cosine similarity metric to quantify the similarity between any two genes. Clustering was performed using R (http://www.r-project.org/).

Chromatin Immunoprecipitation and Library preparation for Next Generation Sequencing:

MESP1 and Histones ChIP:

H3K4me3 ChIP was performed in purified day five UH3 YFP+ cells. MESP1 ChIP was performed in E14 mES cells. E14 cells have wild-type MESP1 alleles and higher MESP1 expression, which produced consistently high-quality ChIP in pilot experiments.

Embryoid bodies (EBs) were collected on day four for MESP1 ChIP-seq. Briefly, the collected EBs was treated with formaldehyde for chromatin cross-linking. Chromatin DNA of the treated cells was fragmented by sonication. The DNA was run on a 1% agarose gel in 2 µL and 4 µL aliquots to check the size of the fragments. Then 1 mL of the nuclear extract from about $1x10^6$ (starting material for hanging drops) cells was used for immunoprecipitations. Portion of the fragmented DNA supernatant was saved for input and the remainder was incubated with MESP1-C antibody H3K27me3 (diagenode), H2AK119ubq (digenode), and EZH2 (diagenode). After overnight incubation, the antibody/DNA complex was immunoprecipitated down using Protein G Agarose beads. Following extensive washing, DNA/protein was eluted, and DNA was separated from the protein complex through digestion with proteinase K and purification using diagenode iPURE kit. The polyclonal antibody against the C-terminus of mouse MESP1 was produced in rabbit using peptide immunogen SSDMLALLETWTPPQEWPPA. It was purified by affinity chromatography, and validated by enzyme-linked immunosorbent assay (ELISA), western blot (Figure S11A), and immunoprecipitation (Figure S11B). Similar procedures were followed for H3K4me3 ChIP, while samples were from sorted day five UH3 YFP+ cells, and the antibody was purchased from cell signaling (9727s).

70 ng of the ChIPed DNA or input DNA was used for Illumina sequencing library generation. First, end repair of ChIP DNA was carried out using T4 DNA polymerase, T4 polynucleotide kinase, and DNA pol I Klenow fragment. Then a dA tail at the 3' end followed by ligation with illumines Tru-seq index adaptors. The

ligated fragments were amplified using limited cycles of PCR. The size and distribution of the amplified library was measured by agarose gel for quality control. The amplified DNA was quantified with the Quant-iT PicoGreen dsDNA Kit (Invitrogen) and diluted to 10 nM. The constructed libraries were sequenced with Solexa sequencing technology at the University of Houston using Illumina's high-throughput Genome Analyzer. The input containing the entire genome was used as a control, and each library was run using a single lane.

Numerous times were attempted to pull out MESP1-bound chromatin using MESP1-Cre;Rosa26-YFP mES line (UH3) using ChIP-Seq. However, the data did not show enrichment indicating few MESP1-bound DNA fragments and hence did not passed the ChIP-Seq data standards laid out by the ENCODE project. This is likely due to the heterozygous MESP1 allele or other intrinsic cell line variations. The E14 line had the same genetic background, and we were able to generate high-quality data from these cells. Thus, the timing was carefully selected for sample collection when MESP1 was expressed at the highest level.

ChIP-Seq data processing: All the ChIP-Seq data were aligned to the mouse reference genome using bowtie with program options: *-t -v 2 -a -m 1 --best --strata*. Reads, which mapped to multiple locations in the genome, were discarded before subsequent analysis.

Histone modification patterns display broad peaks. Therefore, we used MACS2 (https://github.com/taoliu/MACS/) with the "--broad" option to process the H3K4me3 ChIP-Seq data. A stringent FDR threshold of 0.01 was used to obtain the final set of peaks. To compute the H3K4me3 signal or enrichment for a gene,

we considered an interval that flanks 2000 bp upstream and downstream of the transcription start site of the gene. The read depth at every position within the intended interval was computed. The sum of the read depth across the interval was the H3K4me3 signal.

Downstream analysis of ChIP-Seq data: To perform downstream analysis of ChIP-Seq data, we used HOMER to assign gene-specific information (nearest gene, distance from nearest TSS) to each peak based on the peak's nearest annotated TSS. We also classified peaks into promoter (-1kb to +1kb within the TSS), intergenic, intron, and exon. The sequences representing the peaks were extracted from the mouse genome. The background set of sequences (containing matching GC contents to the peaks) generated by HOMER was used to compare the original peak sequences. The binary response, *y*, indicating a motif is present (y=1) or absent (y=0) in a peak was posed as a regression problem: $y \approx a + xb + yc$, where *x* indicates the frequency of the motif in the peak, y is the GC content, and a is constant. We used FIMO to compute the frequency of motifs from JASPAR database in the peaks using a *P*-value threshold of 0.0001. FIMO was applied to both the sequences representing the actual peaks and the background set. Regression analysis using motif counts and GC content in the actual peak sequences and the background set yielded two metrics: *P*-values and z-values. Motifs were ranked according to z-values; starting from the one with the highest z-score. This same procedure was used to compute enrichment for different variants of ebox in the MESP1 peaks.

To identify direct targets, we tested for differential expression from the RNA-Seq data using edgeR (p-value < 0.05).

Enriched GO terms were obtained using DAVID on the set of genes which are nearest to the peaks and using mouse genome as the background set of genes.M.

H3K27me3 Histone modification ChIP-Seq fastq data for GSE47950 was obtained from the GNomEx database under accession numbers 44R and can be found: https://b2b.hci.utah.edu/gnomex/ by the guest login. ODS

ESC Culture and Differentiation

MESP1$^{Cre/1}$; Rosa26$^{EYFP/+}$ reporter (UH3) cells were established by the conventional blastocyst outgrowth method. AB2.2 ESCs were used as a control. Cells were cultured in knockout DMEM (GIBCO, NY, http://www.lifetechnologies.com/us/en/ home/brands/gibco.html) with 15% stem-cell grade fetal-bovine serum (Atlanta Biologicals, GA, https://www.atlantabio. com/), 1% antibiotic/antimitotic (GIBCO, NY), 100 IM Bmercaptoethanol, 2 mM l-glutamine (GIBCO, NY), and supplemented with 1000 U/ml of leukemia inhibitory factor (LIF) to maintain pluripotency. The media was changed daily, and passaging of the cells was done every three days. Before differentiating, the UH3 cells were FACs sorted to remove all YFP+ cells generated from low-level spontaneous differentiation. AB2.2 and UH3 cells were differentiated using the hanging drop method. Briefly, 20 ul drops of LIF free ES media containing 400 cells were placed on the lid of a square petri dish. The bottom of the dish was filled with 5 ml

autoclaved water, and the drop-containing lid was inverted over the dish. The embryoid bodies were plated on day five to gelatin-coated dishes. LIF-free ES media was refreshed every three days. For initial differentiation characterization studies, the ES media was supplemented with 10 ng/ml human BMP-4 (Peprotech, NJ, https://www.peprotech.com/).

6. Appendix B Chapter 2 Supplement

Published Work

Soibam, B., Benham, A., Kim, J., Weng, K. C., Yang, L., Xu, X., Robertson, M., Azares, A., Cooney, AJ., Schwartz, R.J., Liu, Y. (2015). Genome-Wide Identification of MESP1 Targets Demonstrates Primary Regulation Over Mesendoderm Gene Activity. *STEM CELLS, 33*(11), 3254-3265.

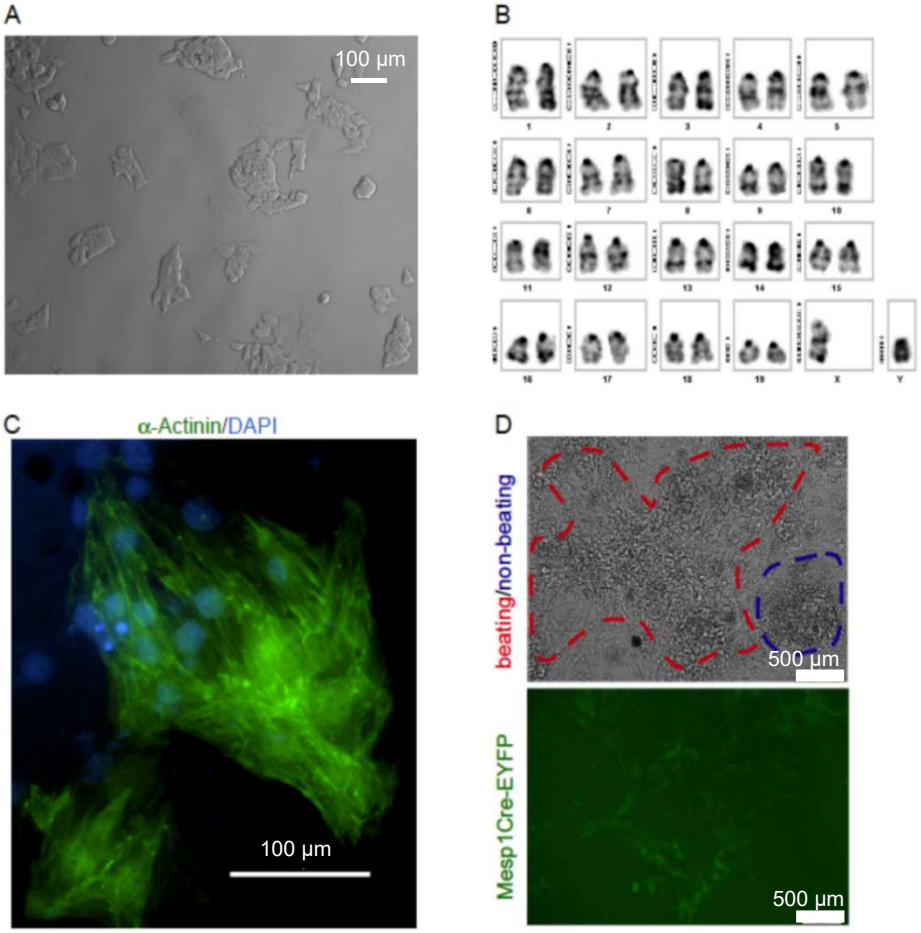


Figure S1: MESP1-lineage cells give rise to beating cardiac myocytes. (A) Phase contrast image of undifferentiated UH3 cells at 100x magnification. (B) Karyotype analysis of UH3 cells showing a normal male karyotype. (C) Representative images of FACs sorted YFP+ UH3 cells staining for well-formed sarcomeres by anti-alpha-actinin antibody. (D) Beating cluster (red circled) co-localized with YFP+ cells signal. The cells within the area circled by a blue line was YFP- and did not participate in spontaneous beating. Scale bars are 500 µm.

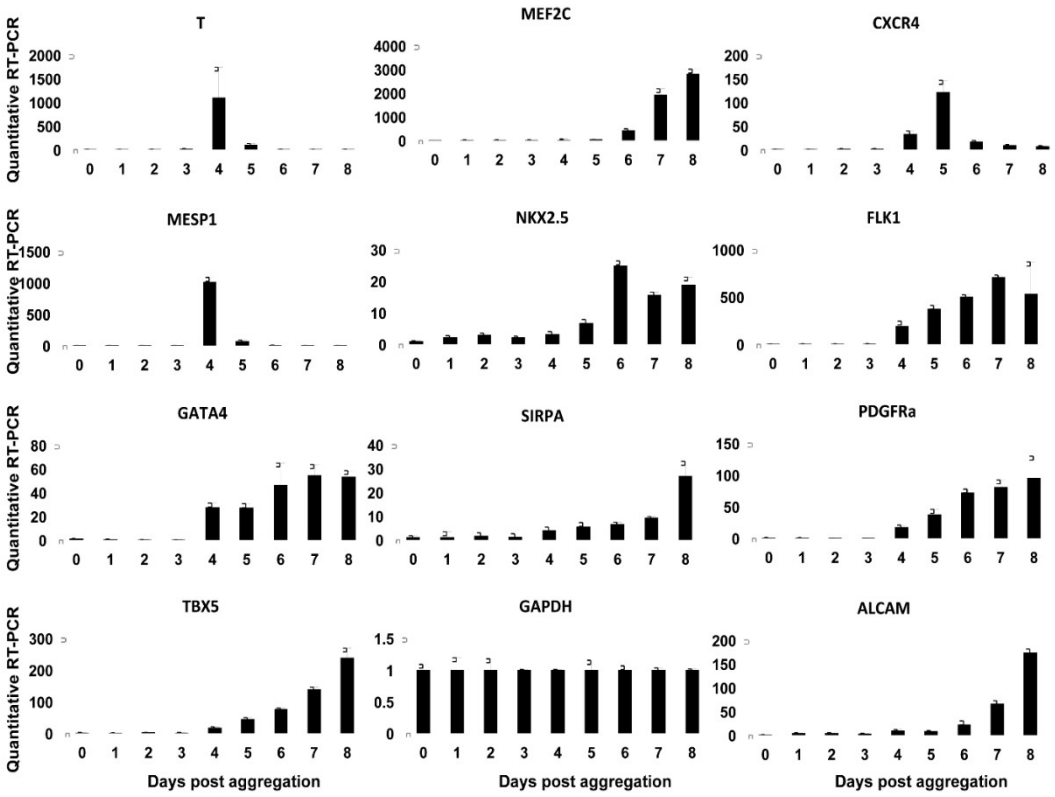


Figure S2. qRT-PCR characterization of lineage-representative genes in UH3 cells. UH3 cells were aggregated and differentiation in hanging drops. At the indicated days (days 0, 1, 2, 3, 4, 5, 6, 7 and 8), EBs were harvested, RNA isolated, and QPCR performed to measure representative gene expression.

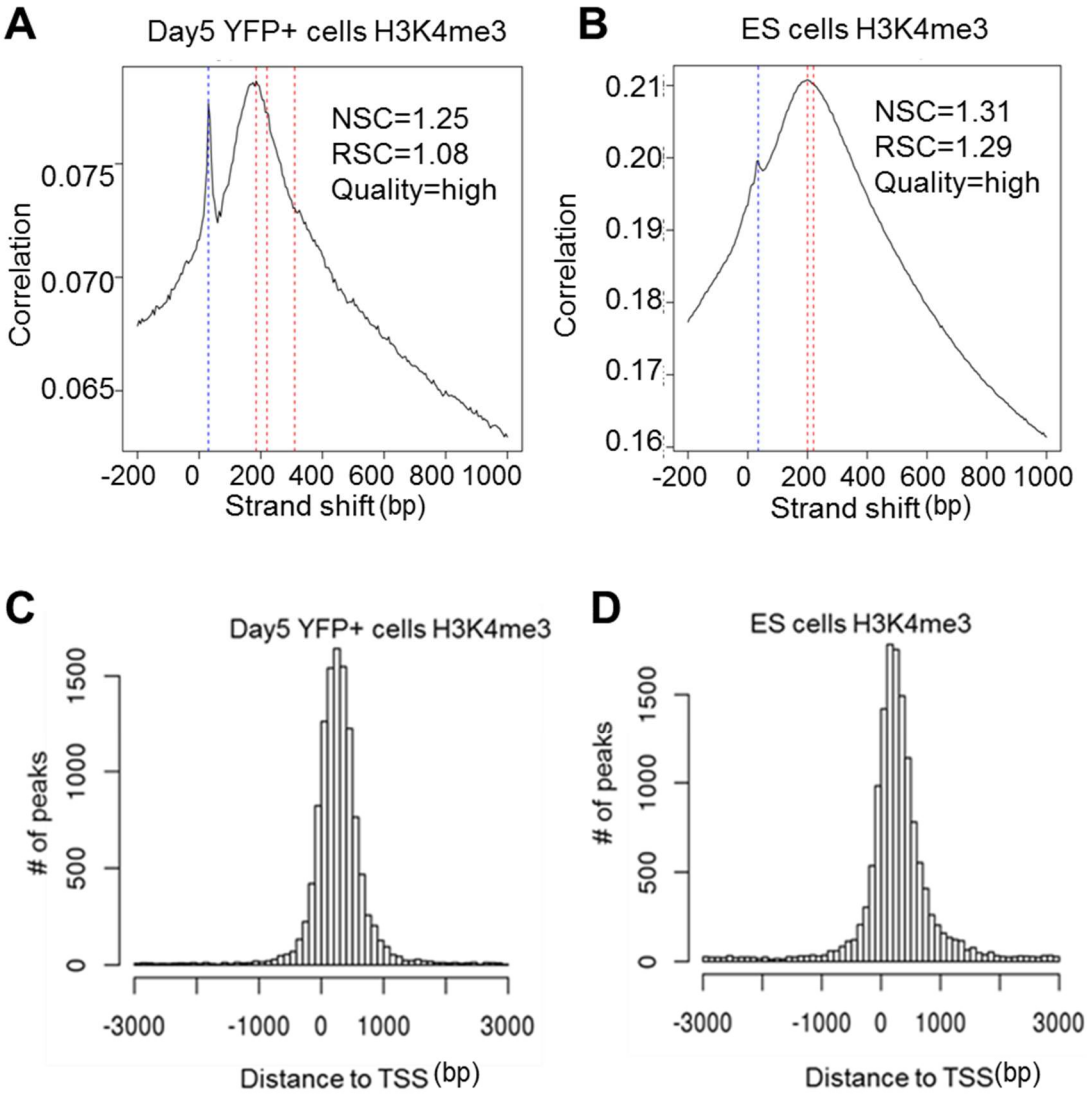


Figure S3. H3K4me3 ChIP-Seq for MESP1-YFP+ cells. (A) and (B) H3K4me3 ChIP-Seq data pass ENCODE ChIP-Seq data quality standard. Cross-correlation curve peaks approximately at 200 bp for day five YFP+ and mES cells. ChIP-Seq quality metrics such as relative strand correlation (RSC), normalized strand correlation (NSC) were above the standard suggested by ENCODE. (C) and (D) Histogram of H3K4me3 peaks with respect to distance from the nearest TSS.

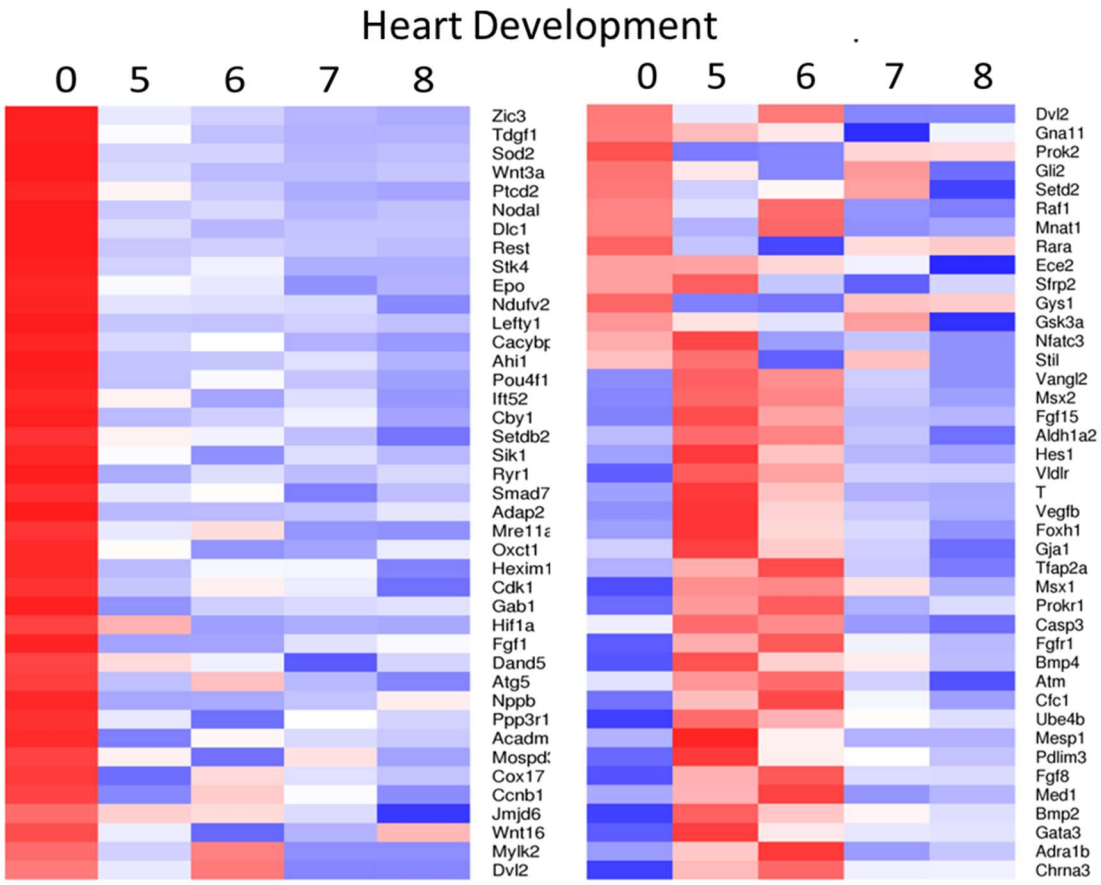


Figure S4: Expanded Heatmaps of genes in main Figure 2E. Expression profiles of RNA-seq (FPKM) data (days 0 (mES cells), 5, 6, 7, 8 post differentiation YFP+ cells) of genes associated with Heart, mesoderm and endoderm development are shown. Heatmaps of z-score of fold changes were split because of large number of genes.

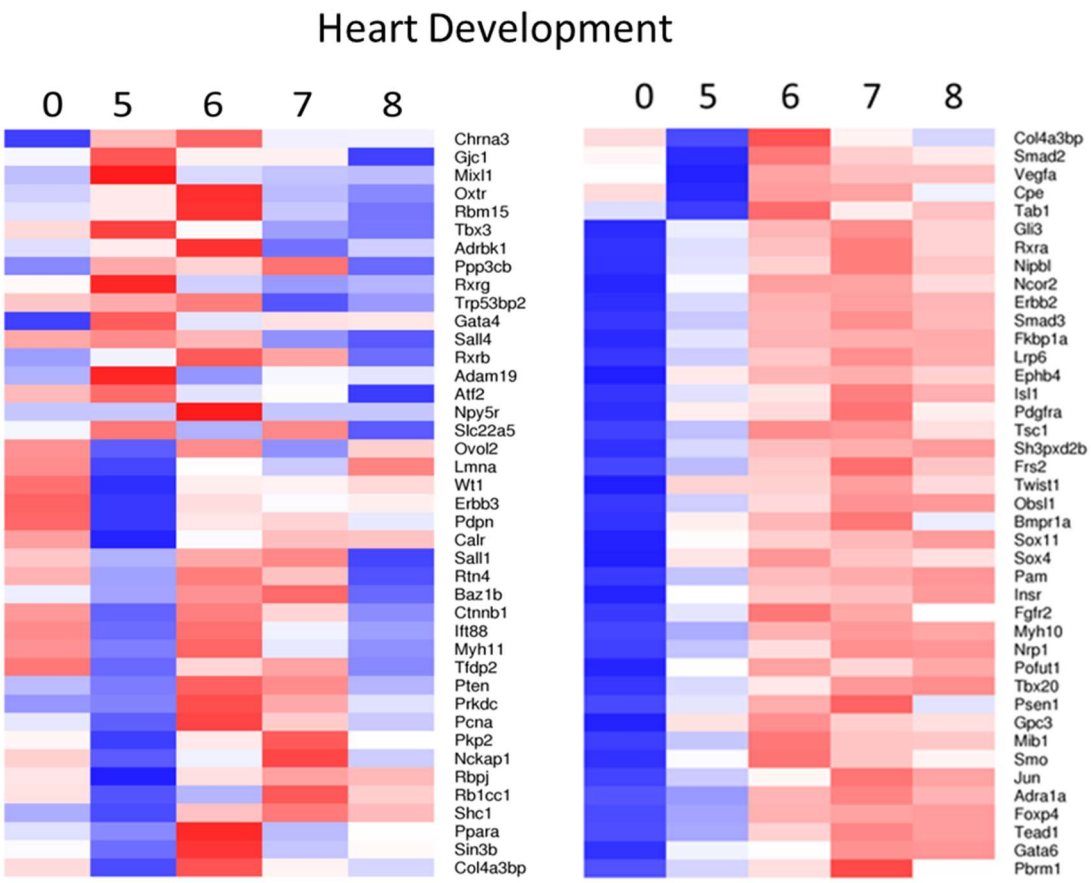


Figure S4 Continued

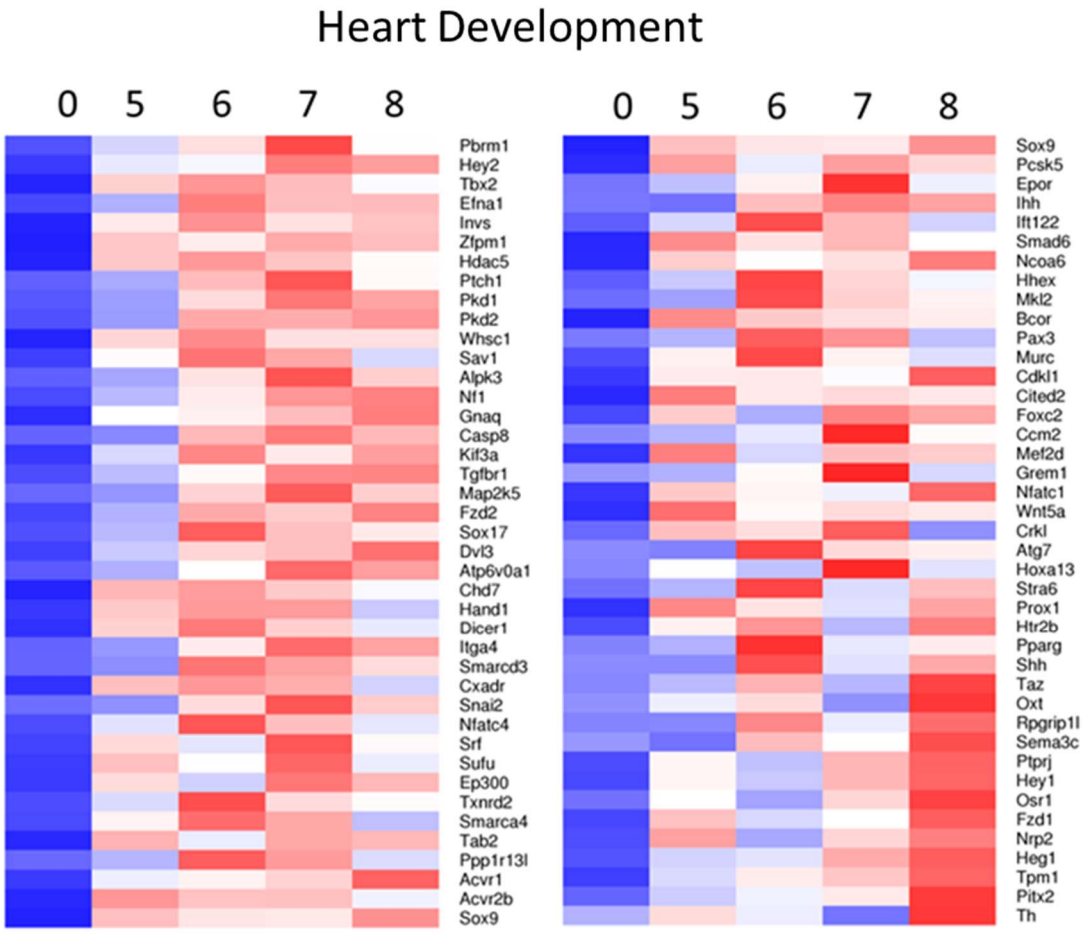


Figure S4 Continued

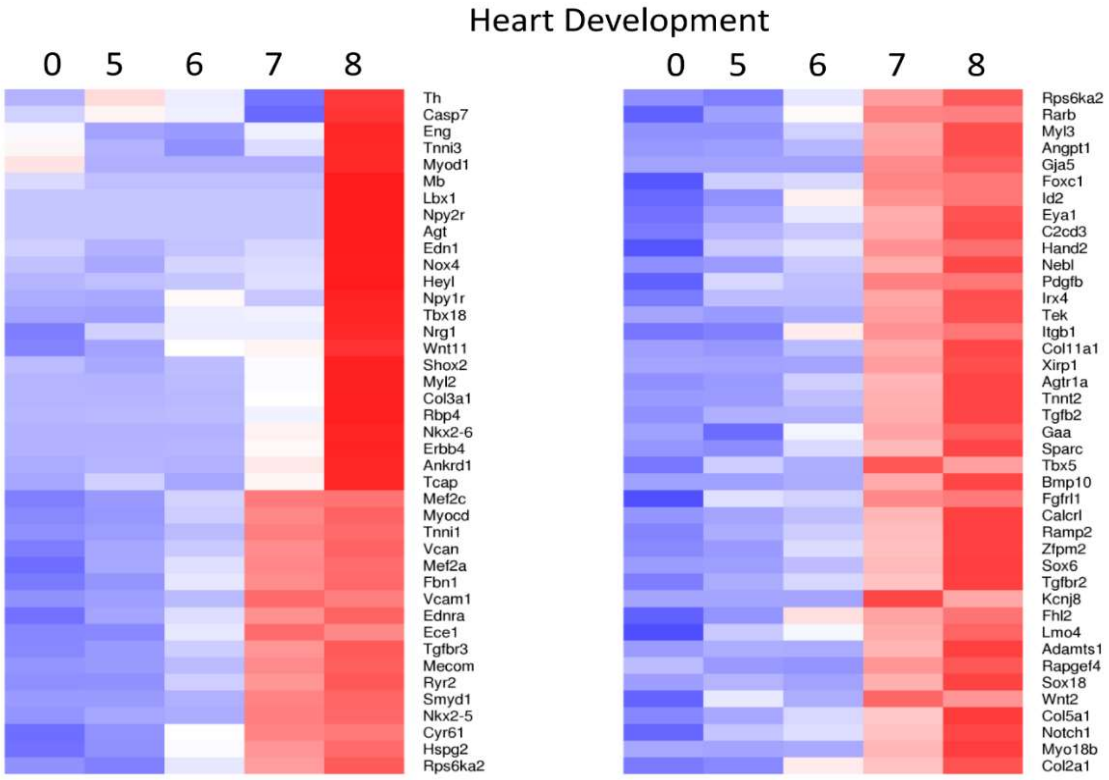


Figure S4 Continued

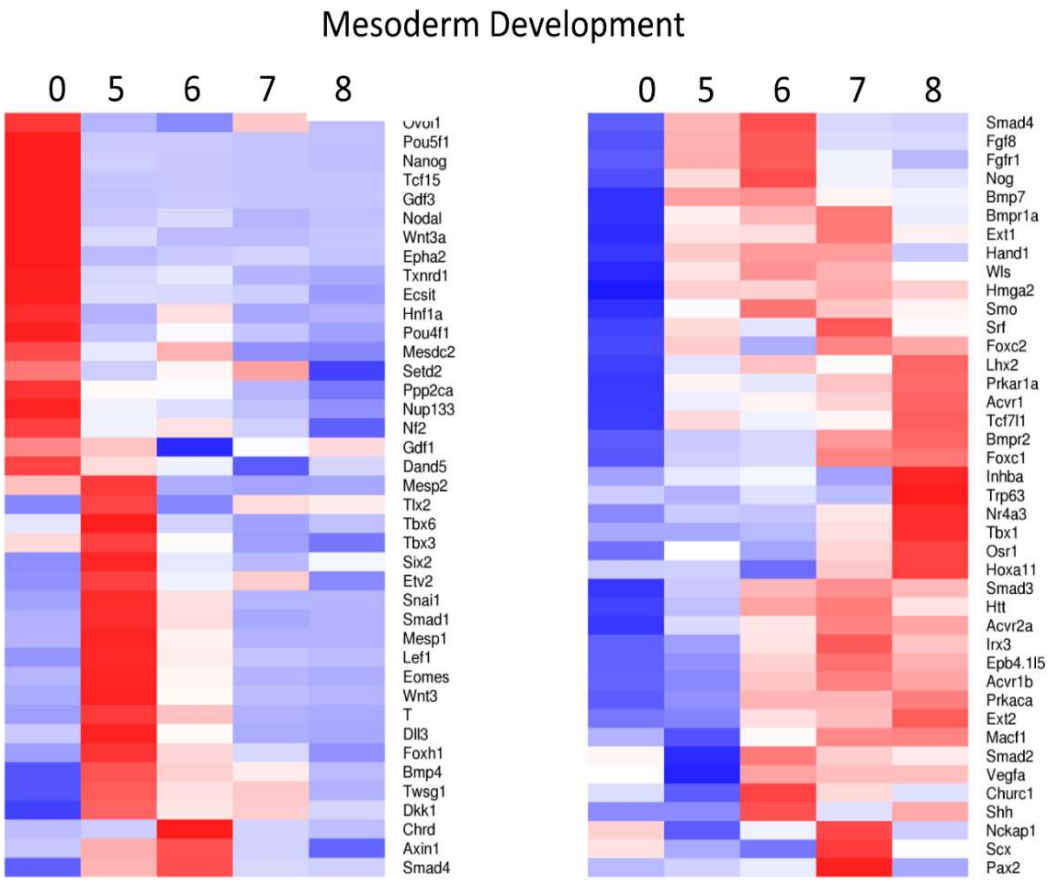


Figure S4 Continued

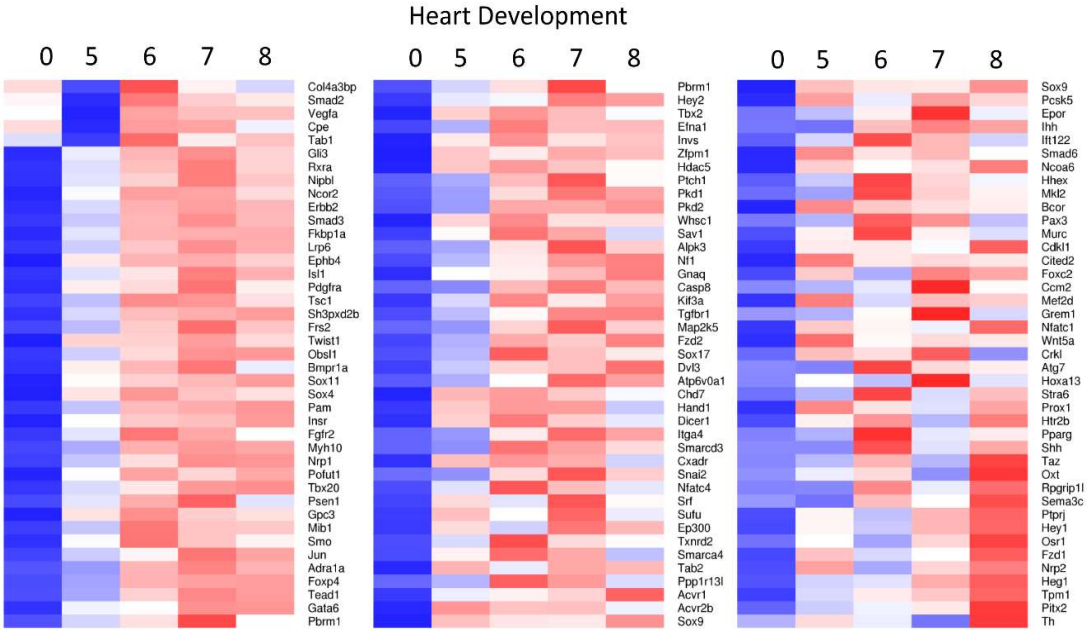


Figure S4 Continued

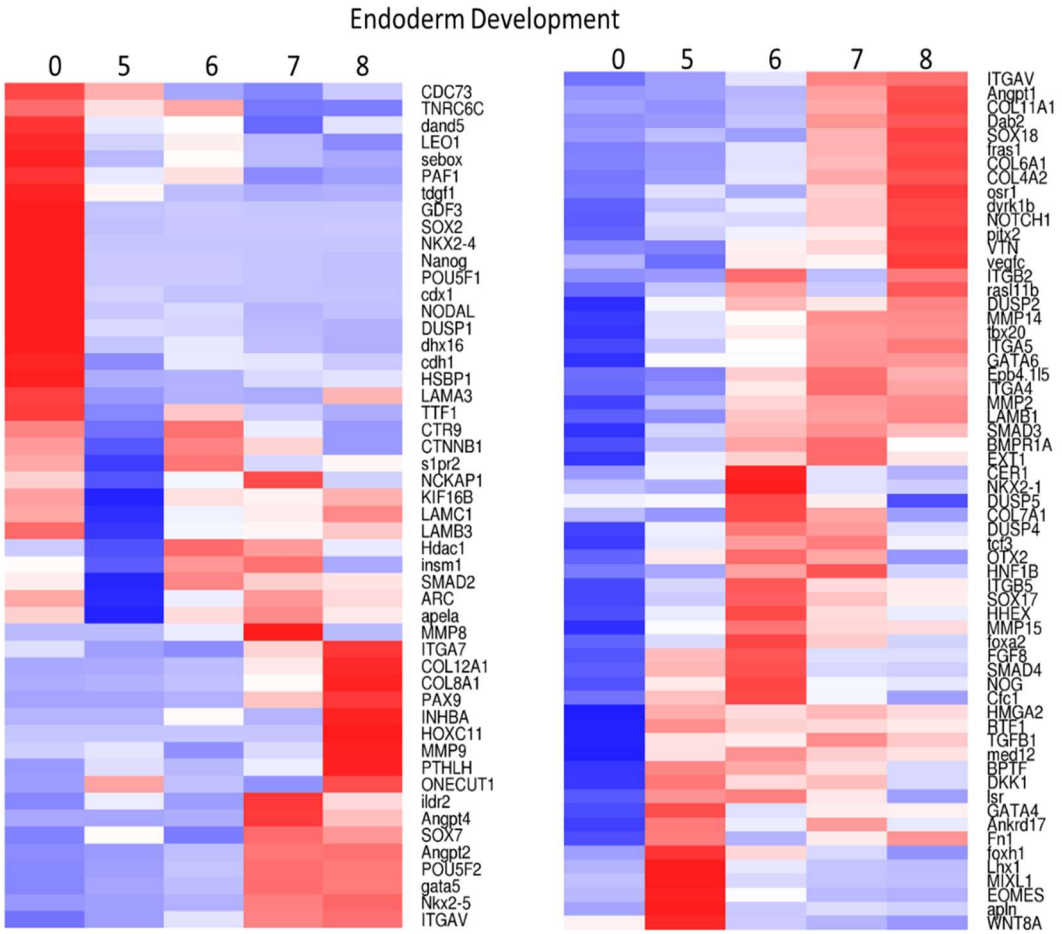


Figure S4 Continued

	Peaks #		%			
Ebox Variant	Mesp1	Background	Mesp1	Background	Score	p-value
CAGCTG	11248	7164	26%	17%	34.3	<2e-16
CACGTG	6749	4131	16%	10%	26.8	<2e-16
CACCTG	9652	6745	22%	16%	25.6	<2e-16
CACATG	3714	2267	9%	5%	21.4	<2e-16
CAGGTG	9572	7802	22%	18%	15.2	<2e-16
CATGTG	3699	2790	9%	6%	13.8	<2e-16
CACTTG	3750	3010	9%	7%	10.7	<2e-16
CAGATG	4072	3353	9%	8%	10.5	<2e-16
CATCTG	4105	3618	9%	8%	7.57	3.67E-14
CAACTG	2554	2160	6%	5%	7.17	7.81E-13
CAAGTG	3566	3152	8%	7%	6.57	4.98E-11
CAGTTG	2536	2341	6%	5%	4.15	3.34E-05
CAAATG	1147	1429	3%	3%	-4.2	2.64E-05
CATTTG	1226	1591	3%	4%	-5.45	5.08E-08
CATATG	538	818	1%	2%	-5.81	6.36E-09
CAATTG	596	917	1%	2%	-7.37	1.74E-13

Figure S5. Ebox variant enrichment in MESP1 peaks. The variants colored in light blue (Note the positive scores) are enriched in MESP1 peaks compared to background. There are some variants, which occupy less than 10% of the MESP1 peaks. These variants might be less preferred than the other variants.

A

Motif	Consensus	P-value	MESP1 peaks	% of peaks	# of Background	%
Tcf12(HLH)	VCAGCTGYTG	1.00E-75	10324	23.82%	7763.3	20.19%
Ap4(HLH)	NAHCAGCTGD	1.00E-74	10627	24.52%	8026.1	20.87%
NF1-halfsite(CTF)	YTGCCAAG	1.00E-72	12922	29.81%	9977.9	25.95%
n-Myc(HLH)	VRCCACGTGG	1.00E-71	5713	13.18%	4020.5	10.46%
Ptf1a(HLH)	ACAGCTGTTN	1.00E-70	20780	47.94%	16794.7	43.68%
Tlx?	CTGGCAGSCTGCCA	1.00E-59	3928	9.06%	2682.8	6.98%
c-Myc	VVCCACGTGG	1.00E-57	4140	9.55%	2863.3	7.45%
Atoh1(bHLH)	VNRVCAGCTGGY	1.00E-51	9124	21.05%	6992.1	18.18%
E2A(HLH)	DNRCAGCTGY	1.00E-51	14533	33.53%	11597.3	30.16%
c-Myc(HLH)	VCCACGTG	1.00E-49	4738	10.93%	3396.2	8.83%

B

MOTIF NAME	# of peaks	# background	% of peaks	% of background	P-Value	Z-score
MA0522.1_Tcf3	13013	8341	30%	19%	2.42E-302	37.17
MA0500.1_Myog	11092	7424	26%	17%	3.41E-204	30.49
MA0521.1_Tcf12	10635	7100	25%	16%	3.80E-197	29.95
MA0461.1_Atoh1	9057	5831	21%	13%	2.13E-193	29.67
MA0499.1_Myod1	11292	7728	26%	18%	1.63E-189	29.36
MA0048.1_NHLH1	10491	7367	24%	17%	5.41E-134	24.63
MA0464.1_Bhlhe40	4849	2897	11%	7%	1.41E-121	23.45
MA0147.2_Myc	3963	2261	9%	5%	5.71E-119	23.19
MA0104.3_Mycn	3495	2031	8%	5%	2.30E-93	20.50
MA0138.2_REST	10190	7793	24%	18%	9.03E-79	18.79

Figure S6. Top 10 motifs in MESP1 peaks identified by HOMER (A) and FIMO (B).

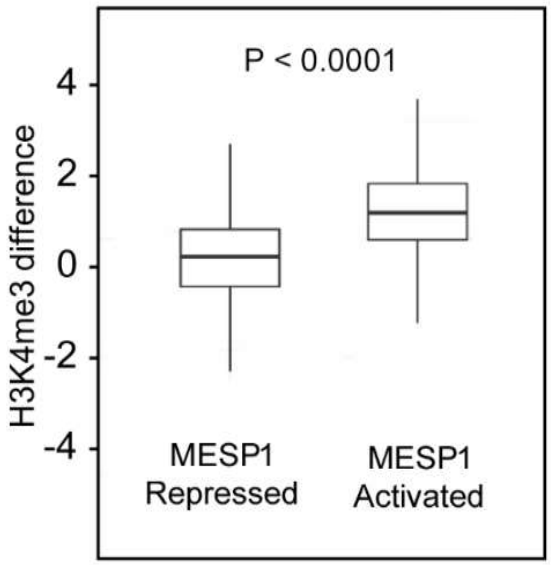


Figure S7. Increase in H3K4me3 modification associated with MESP1-activated targets. We computed the normalized ChIP-Seq reads (both in mES and day five MESP1 YFP+ cells) which overlap with a 4 kb interval centered at the TSS of the MESP1 direct targets. The difference in the H3K4me3 signal was computed by subtracting the H3K4me3 signal in mES cells from that in day five MESP1 YFP+ cells. MESP1-activated targets showed significant increase compared to repressed targets in H3K4me3 signal from mES to day five YFP+ cells.

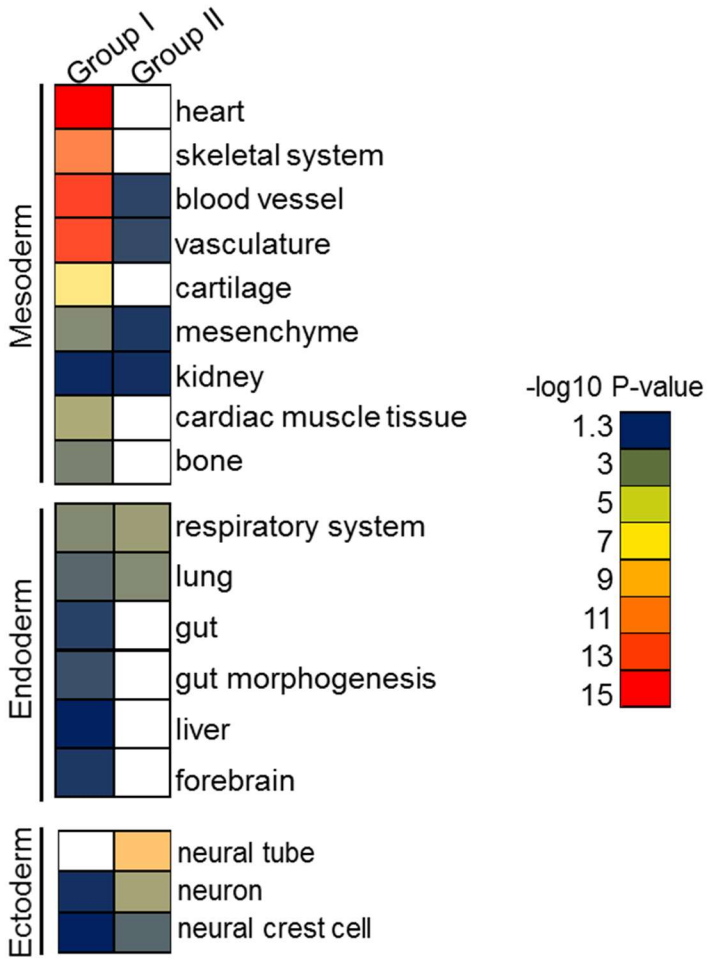


Figure S8. GO enrichment of MESP1-activated targets in YFP+ and YFP- cells. The MESP1-activated targets were compared at different time points (day 5, 6, 7 and 8) for differentially expression between YFP+ and YFP- cells. The heatmap shows enriched gene ontology terms for two groups of MESP1-activated targets. Group I comprises of all targets which were strictly upregulated in YFP+ cells during course of differentiation, group II represents targets which were strictly up-regulated in YFP- cells during the course of differentiation. The color scale is shown in –log10 P-value.

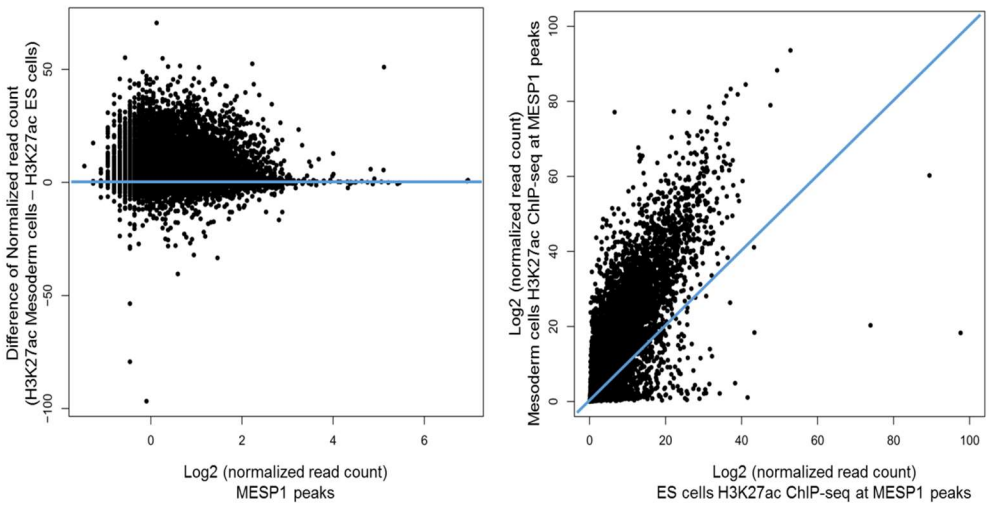


Figure S9. MESP1 induces genome-wide H3K27ac histone modification. (A) Increase in H3K27ac signal at MESP1 peak regions from mES to mesoderm cells. At all the MESP1 peak locations, the difference between H3K27ac ChIP signal in mesoderm and mES cells were plotted as a function of log2 (normalized read count). The horizontal blue line represents zero difference. Majority of the data points lie above the light blue line. (B) At all the MESP1 peak locations, H3K27ac ChIP signal in mesoderm cells and mES cells were plotted. Majority of the data points lie above the light blue line indicating an increase in H3K27ac modifications (increase in mesoderm cells compared to mES cells).

Primer ID from previous study (2)	ChIP-PCR result shown in (2)	Gene	Overlap with this Study MESP1 ChIP-Seq Peaks
cFoxA2AF	x	Foxa2	
cFoxA2BF		Foxa2	
cFoxA2CF		Foxa2	
cFoxA2DF	x	Foxa2	
cG4AF		Gata4	X
cG4BF	x	Gata4	
cG4CF	x	Gata4	
cG4DF	x	Gata4	
cG4EF		Gata4	X
cG4FF		Gata4	
cG4GF		Gata4	
cGSCAF1		Gsc	
cGSCBF1	x	Gsc	
cGSCCF1		Gsc	
cHand2AF1		Hand2	
cHand2BF1		Hand2	X
cHand2CF1		Hand2	
cHand2DF1	x	Hand2	
cHand2EF1	x	Hand2	
cHand2FF1	x	Hand2	
cHand2GF1		Hand2	
cHand2HF1		Hand2	
cHand2IF1		Hand2	
cHand2JF1		Hand2	
cHand2KF1		Hand2	
cHand2LF1	x	Hand2	
cHand2MF1	x	Hand2	

Figure S10. Comparison of MESP1 DNA binding regions identified by ChIP-PCR and ChIP-Seq. We compiled a list of DNA regions which were tested for MESP1 enrichment using ChIP-PCR as described in a previous study. The primers representing these regions are shown in the first column of the table. An ‘x’ in the second column indicates positive results in ChIP-PCRs. In the fourth column of the table, an ‘x’ indicates an overlap (at most at a distance of 500 bp) with an MESP1 ChIP-Seq peak in this study. About 25% of the MESP1 enriched regions identified in the published ChIP-PCR were detected by our MESP1 ChIP-Seq data.

cM1AF		MESP1	X
cM1BF	x	MESP1	
cM1CF		MESP1	
cM1DF		MESP1	
cMyocAF		Myocd	
cMyocBF	x	Myocd	X
cMyocBF	x	Myocd	X
cMyocCF	x	Myocd	X
cMyocCOF		Myocd	
cN24AF	x	NKx2-5	X
cN25BF		Nkx2-5	X
cN25C		NKx2-5	
cN25D	x	NKx2-5	X
cN25E	x	NKx2-5	X
cRply2AF	x	RIPPLY2	
cRply2BF		RIPPLY2	
cRply2CF	x	RIPPLY2	
cRply2COF	x	RIPPLY2	
cSox17AF2		Sox17	
cSox17BF2		Sox17	
cSox17cF2	x	Sox17	
cSox17DF2	x	Sox17	
cSox17EF2		Sox17	
cBrachAF	x	T	X
cBrachBF		T	X
cBrachCF	x	T	
cDkk1AF	x	Dkk1	
cDKK1BF	x	Dkk1	

Figure S10. Cont.

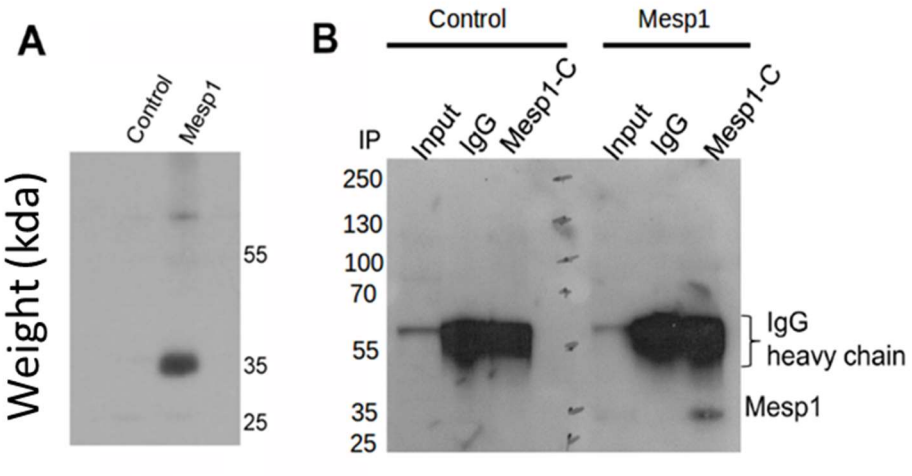


Figure S11. MESP1 antibody validation. (A) MESP1-C antibody detected a prominent band of predicted mass in western blot analysis. Lysate was prepared in 293T cells expressing ectopic mouse MESP1. (B) MESP1-C antibody can be used to immunoprecipitate MESP1. Lysate was prepared from mES cells fixed with 1% PFA to mimic ChIP conditions and expressing ectopic mouse MESP1. MESP1-C antibody was used in IP followed by western blot.

7. Appendix C Chapter 3 Supplement

MESP1 Represses the Neuro-ectoderm Program by recruiting the PRC2 complex to specific ebox Variant

Figure S1. List of GO terms from heatmap in chapter 3 Figure 3.2.

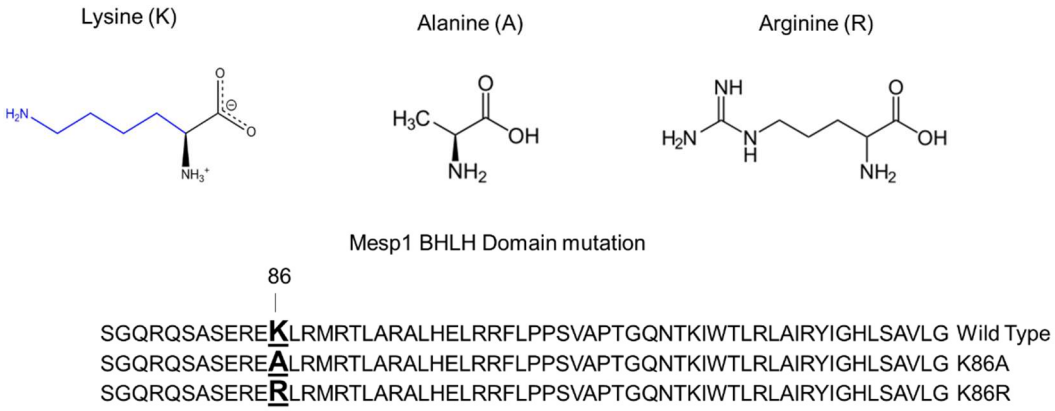


Figure S2. Diagram of Mesp1 mutation and location. Lysine 86 is located in the conserved E/(RK)N(R/N) motif. Mutation into alanine eliminates interaction with phosphate back bone. Mutation into Arginine potentially preserve interaction but change structural positioning

8. Reference

Anders, S., S. Anders and W. Huber (2010). "Differential expression analysis for sequence count data." Nature Precedings **11**(10).

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, M. J. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin and G. Sherlock (2000). "Gene Ontology: tool for the unification of biology." Nature Genetics **25**(1): 25-29.

Atchley, W. R. and W. M. Fitch (1997). "A natural classification of the basic helix-loop-helix class of transcription factors." Proceedings of the National Academy of Sciences of the United States of America **94**(10): 5172-5176.

Bannister, A. J. and T. Kouzarides (2011). "Regulation of chromatin by histone modifications." Cell Research **21**(3): 381-395.

Blackledge, N. P., N. R. Rose and R. J. Klose (2015). "Targeting Polycomb systems to regulate gene expression: modifications to a complex story." Nature reviews. Molecular cell biology **16**(11): 643-649.

Bondue, A. and C. Blanpain (2010). "Mesp1 A Key Regulator of Cardiovascular Lineage Commitment." Circulation Research **107**(12): 1414-1427.

Bondue, A., G. Lapouge, C. Paulissen, C. Semeraro, M. Iacovino, M. Kyba and C. Blanpain (2008). "Mesp1 Acts as a Master Regulator of Multipotent Cardiovascular Progenitor Specification." Cell Stem Cell **3**(1): 69-84.

Bondue, A., S. Tännler, G. Chiapparo, S. Chabab, M. Ramialison, C. Paulissen, B. Beck, R. Harvey and C. Blanpain (2011). "Defining the earliest step of cardiovascular progenitor specification during embryonic stem cell differentiation." The Journal of Cell Biology **192**(5): 751-765.

Cao, Y., Z. Yao, D. Sarkar, M. Lawrence, G. J. Sanchez, M. H. Parker, K. L. MacQuarrie, J. Davison, M. T. Morgan, W. L. Ruzzo, R. C. Gentleman and S. J. Tapscott (2009). "Genome-wide MyoD Binding in Skeletal Muscle Cells: A Potential for Broad Cellular Reprogramming." Developmental Cell **18**(4): 662-674.

Carver, E. A., R. Jiang, Y. Lan, K. F. Oram and T. Gridley (2001). "The Mouse Snail Gene Encodes a Key Regulator of the Epithelial-Mesenchymal Transition." Molecular and Cellular Biology **21**(23): 8184-8188.

Chan, S., X. Shi, A. Toyama, R. W. Arpke, A. Dandapat, M. Iacovino, J. Kang, G. Le, H. R. Hagen, D. J. Garry and M. Kyba (2013). "Mesp1 Patterns Mesoderm into Cardiac, Hematopoietic, or Skeletal Myogenic Progenitors in a Context-Dependent Manner." Cell Stem Cell **12**(5): 587-601.

Chiapparo, G., X. Lin, F. Lescroart, S. Chabab, C. Paulissen, L. Pitisci, A. Bondue and C. Blanpain (2016). "Mesp1 controls the speed, polarity, and directionality of cardiovascular progenitor migration." The Journal of Cell Biology **213**(4): 463-477.

Cho, J.-H. and M.-J. Tsai (2004). "The Role of BETA2/NeuroD1 in the Development of the Nervous System." Molecular Neurobiology **30**(1): 35-48.

Costello, I., I.-M. Pimeisl, S. Dräger, E. K. Bikoff, E. J. Robertson and S. J. Arnold (2011). "The T-box transcription factor Eomesodermin acts upstream of Mesp1 to specify cardiac mesoderm during mouse gastrulation." Nature Cell Biology **13**(9): 1084-1091.

David, R., V. Jarsch, F. Schwarz, P. Nathan, M. Gegg, H. Lickert and W.-M. Franz (2011). "Induction of MesP1 by Brachyury(T) generates the common multipotent cardiovascular stem cell." Cardiovascular Research **92**(1): 115-122.

David, R., F. Schwarz, C. Rimmbach, P. Nathan, J. Jung, C. Brenner, V. Jarsch, J. Stieber and W.-M. Franz (2013). "Selection of a common multipotent cardiovascular stem cell using the 3.4-kb MesP1 promoter fragment." Basic Research in Cardiology **108**(1): 312.

Devine, W. P., J. D. Wythe, M. George, K. Koshiba-Takeuchi and B. G. Bruneau (2014). "Early patterning and specification of cardiac progenitors in gastrulating mesoderm." eLife **3**.

Dubois, N. C., A. M. Craft, P. Sharma, D. A. Elliott, E. G. Stanley, A. G. Elefanty, A. Gramolini and G. Keller (2011). "SIRPA is a specific cell-surface marker for isolating cardiomyocytes derived from human pluripotent stem cells." Nature Biotechnology **29**(11): 1011-1018.

Fiszbein, A., L. E. Giono, A. Quaglino, B. G. Berardino, L. Sigaut, C. von Bilderling, I. E. Schor, J. H. Steinberg, M. Rossi, L. Pietrasanta, II, J. J. Caramelo, A. Srebrow and A. R. Kornblihtt (2016). "Alternative Splicing of G9a Regulates Neuronal Differentiation." Cell reports **14**(12): 2797-2808.
Franklin, A., A. Kao, S. Tapscott and A. Unis (2001). "NeuroD homologue expression during cortical development in the human brain." Journal of child neurology **16**(11): 849-853.

Fu, J.-D., N. R. Stone, L. Liu, I. C. Spencer, L. Qian, Y. Hayashi, P. Delgado-Olguin, S. Ding, B. G. Bruneau and D. Srivastava (2013). "Direct Reprogramming of Human Fibroblasts toward a Cardiomyocyte-like State." Stem Cell Reports **1**(3): 235-247.

Gao, Z., K. Ure, J. L. Ables, D. C. Lagace, K.-A. Nave, S. Goebbels, A. J. Eisch and J. Hsieh (2009). "Neurod1 is essential for the survival and maturation of adult-born neurons." Nature Neuroscience **12**(9): 1090-1092.

Goessling, W. and D. Y. Stainier (2016). "Chapter 14 Endoderm specification and liver development." Methods in Cell Biology **134**: 463-483.

Goodwin, S., J. D. McPherson and R. W. McCombie (2016). "Coming of age: ten years of next-generation sequencing technologies." Nature Reviews Genetics **17**(6): 333-351.

Grant, C. E., T. L. Bailey and W. Noble (2011). "FIMO: scanning for occurrences of a given motif." Bioinformatics **27**(7): 1017-1018.

Guo, Y., S. Liu, X. Zhang, L. Wang, X. Zhang, A. Hao, A. Han and J. Yang (2014). "Sox11 promotes endogenous neurogenesis and locomotor recovery in mice spinal cord injury." Biochemical and Biophysical Research Communications **446**(4): 830-835.

Haraguchi, S., S. Kitajima, A. Takagi, H. Takeda, T. Inoue and Y. Saga (2001). "Transcriptional regulation of Mesp1 and Mesp2 genes: differential usage of enhancers during development." Mechanisms of Development **108**(1-2): 59-69.

Heinz, S., C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh and C. K. Glass (2010). "Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities." Molecular Cell **38**(4): 576-589.

Hogan, B. and K. S. Zaret (2002). "15 Development of the Endoderm and Its Tissue Derivatives." Mouse Development: 301-330.

Huang, D. W., B. T. Sherman and R. A. Lempicki (2008). "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources." Nat. Protocols **4**(1): 44-57.

Huang, D. W., B. T. Sherman and R. A. Lempicki (2009). "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists." Nucleic Acids Research **37**(1): 1-13.

Islas, J., Y. Liu, K.-C. Weng, M. J. Robertson, S. Zhang, A. Prejusa, J. Harger, D. Tikhomirova, M. Chopra, D. Iyer, M. Mercola, R. G. Oshima, J. T. Willerson, V. N. Potaman and R. J. Schwartz (2012). "Transcription factors ETS2 and MESP1 transdifferentiate human dermal fibroblasts into cardiac progenitors." Proceedings of the National Academy of Sciences **109**(32): 13016-13021.

Jones, S. (2004). "An overview of the basic helix-loop-helix proteins." Genome biology **5**(6): 226.

Kan, L., N. Israsena, Z. Zhang, M. Hu, L.-R. Zhao, A. Jalali, V. Sahni and J. A. Kessler (2004). "Sox1 acts through multiple independent pathways to promote neurogenesis." Developmental Biology **269**(2): 580-594.

Kele, J., N. Simplicio, A. L. M. Ferri, H. Mira, F. Guillemot, E. Arenas and S.-L. Ang (2006). "Neurogenin 2 is required for the development of ventral midbrain dopaminergic neurons." Development (Cambridge, England) **133**(3): 495-505.

Kennison, J. (2002). "Encyclopedia of Molecular Biology." wiley.

Khan, A., A. Lee and T.-Y. Roh (2015). "Polycomb group protein-mediated histone modifications during cell differentiation." Epigenomics **7**(1): 75-84.

Kim, D., G. Pertea, C. Trapnell, H. Pimentel, R. Kelley and S. L. Salzberg (2013). "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions." Genome Biology **14**(4): 1-13.

Kimelman, D. and K. J. P. Griffin (2000). "Vertebrate mesendoderm induction and patterning." Current Opinion in Genetics & Development **10**(4): 350-356.

Kinder, S. J., T. E. Tsang, M. Wakamiya, H. Sasaki, R. R. Behringer, A. Nagy and P. P. Tam (2001). "The organizer of the mouse gastrula is composed of a dynamic population of progenitor cells for the axial mesoderm." Development (Cambridge, England) **128**(18): 3623-3634.

Kitajima, S., S. Miyagawa-Tomita, T. Inoue, J. Kanno and Y. Saga (2006). "Mesp1-nonexpressing cells contribute to the ventricular cardiac conduction system." Developmental Dynamics **235**(2): 395-402.

Kitajima, S., A. Takagi, T. Inoue and Y. Saga (2000). "MesP1 and MesP2 are essential for the development of cardiac mesoderm." Development (Cambridge, England) **127**(15): 3215-3226.

Kyba, M. (2016). "Mesoderm, Cooked Up Fast and Served to Order." Cell stem cell **19**(2): 146-148.
Landt, S. G., G. K. Marinov, A. Kundaje, P. Kheradpour, F. Pauli, S. Batzoglou, B. E. Bernstein, P. Bickel, J. B. Brown, P. Cayting, Y. Chen, G. DeSalvo, C. Epstein, K. I. Fisher-Aylor, G. Euskirchen, M. Gerstein, J. Gertz, A. J. Hartemink, M. M. Hoffman, V. R. Iyer, Y. L. Jung, S. Karmakar, M. Kellis, P. V. Kharchenko, Q. Li, T. Liu, S. X. Liu, L. Ma, A. Milosavljevic, R. M. Myers, P. J. Park, M. J. Pazin, M. D. Perry, D. Raha, T. E. Reddy, J. Rozowsky, N. Shoresh, A. Sidow, M. Slattery, J. A. Stamatoyannopoulos, M. Y. Tolstorukov, K. P. White, S. Xi, P. J. Farnham, J. D. Lieb, B. J. Wold and M. Snyder (2012). "ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia." Genome research **22**(9): 1813-1831.

Langmead, B., C. Trapnell, M. Pop and S. L. Salzberg (2009). "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome." Genome Biology **10**(3): 1-10.

Ledent, V. and M. Vervoort (2001). "The Basic Helix-Loop-Helix Protein Family: Comparative Genomics and Phylogenetic Analysis." Genome Research **11**(5): 754-770.

Lee, J. E. (1997). "Basic helix-loop-helix genes in neural development." Current Opinion in Neurobiology **7**(1): 13-20.

Lescroart, F., S. Chabab, X. Lin, S. Rulands, C. Paulissen, A. Rodolosse, H. Auer, Y. Achouri, C. Dubois, A. Bondue, B. D. Simons and C. Blanpain (2014). "Early lineage restriction in temporally distinct populations of Mesp1 progenitors during mammalian heart development." Nature Cell Biology **16**(9): 829-840.

Li, Y., W. Yu, A. J. Cooney, R. J. Schwartz and Y. Liu (2013). "Brief report: Oct4 and canonical Wnt signaling regulate the cardiac lineage factor Mesp1 through a Tcf/Lef-Oct4 composite element." Stem cells (Dayton, Ohio) **31**(6): 1213-1217.

Lim, J.-W. and K. L. Kroll (2008). "Roles of Neurogenin and Geminin in vertebrate neurogenesis." Developmental Biology **319**(2): 486.

Lindsley, C. R., J. G. Gill, T. L. Murphy, E. M. Langer, M. Cai, M. Mashayekhi, W. Wang, N. Niwa, J. M. Nerbonne, M. Kyba and K. M. Murphy (2007). "Mesp1 Coordinately Regulates Cardiovascular Fate Restriction and Epithelial-Mesenchymal Transition in Differentiating ESCs." Cell Stem Cell **3**(1): 55-68.

Liu, E. T. (2005). "Systems Biology, Integrative Biology, Predictive Biology." Cell **121**(4): 505-506.
Liu, Y., M. Asakura, H. Inoue, T. Nakamura, M. Sano, Z. Niu, M. Chen, R. J. Schwartz and M. D. Schneider (2007). "Sox17 is essential for the specification of cardiac mesoderm in embryonic stem cells." Proceedings of the National Academy of Sciences **104**(10): 3859-3864.

Liu, Y., L. Chen, A. Diaz, A. Benham, X. Xu, C. S. Wijaya, F. Fa'ak, W. Luo, B. Soibam, A. Azares, W. Yu, Q. Lyu, D. M. Stewart, P. Gunaratne, A. Cooney, B. K. McConnell and R. J. Schwartz (2016). "Mesp1 Marked Cardiac Progenitor Cells Repair Infarcted Mouse Hearts." Scientific Reports **6**: 31457.

Liu, Y., R. Kaneda, T. W. Leja, T. Subkhankulova, O. Tolmachov, G. Minchiotti, R. J. Schwartz, M. Barahona and M. D. Schneider (2014). "Hhex and Cer1 Mediate the Sox17 Pathway for Cardiac Mesoderm Formation in Embryonic Stem Cells." STEM CELLS **32**(6): 1515-1526.

Ma, Q., C. Fode, F. Guillemot and D. J. Anderson (1999). "Neurogenin1 and neurogenin2 control two distinct waves of neurogenesis in developing dorsal root ganglia." Genes & development **13**(13): 1717-1728.

Ma, Q., C. Kintner and D. J. Anderson (1996). "Identification of neurogenin, a Vertebrate Neuronal Determination Gene." Cell **87**(1): 43-52.

Ma, W., R.-T. Yan, W. Mao and S.-Z. Wang (2009). "Neurogenin3 promotes early retinal neurogenesis." Molecular and Cellular Neuroscience **40**(2): 187-198.

Margueron, R. and D. Reinberg (2011). "The Polycomb complex PRC2 and its mark in life." Nature **469**(7330): 343-349.

Masi, F., C. A. Grove, A. Vedenko, A. Alibés, S. S. Gisselbrecht, L. Serrano, M. L. Bulyk and A. J. M. Walhout (2011). "Using a structural and logics systems approach to infer bHLH–DNA binding specificity determinants." Nucleic Acids Research **39**(11): 4553-4563.

Mathelier, A., X. Zhao, A. W. Zhang, F. Parcy, R. Worsley-Hunt, D. J. Arenillas, S. Buchman, C.-y. Chen, A. Chou, H. Ienasescu, J. Lim, C. Shyr, G. Tan, M. Zhou, B. Lenhard, A. Sandelin and W. W. Wasserman (2014). "JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles." Nucleic Acids Research **42**(D1).

Morimoto, M., M. Kiso, N. Sasaki and Y. Saga (2006). "Cooperative Mesp activity is required for normal somitogenesis along the anterior-posterior axis." Developmental biology **300**(2): 687-698.

Nelson, T. J., R. S. Faustino, A. Chiriac, R. Crespo-Diaz, A. Behfar and A. Terzic (2008). "CXCR4+/FLK-1+ Biomarkers Select a Cardiopoietic Lineage from Embryonic Stem Cells." STEM CELLS **26**(6): 1464-1473.

Oginuma, M., T. Hirata and Y. Saga (2007). "Identification of presomitic mesoderm (PSM)-specific Mesp1 enhancer and generation of a PSM-specific Mesp1/Mesp2-null mouse using BAC-based rescue technology." Mechanisms of Development **125**(5-6): 432-440.

Ohtsuka, T. and R. Kageyama (2010). "The Basic Helix-Loop-Helix Transcription Factors in Neural Differentiation." springer: 15-34.

Papaioannou, V. E. (2004). "24 Early Embryonic Mesoderm Development." Handbook of Stem Cells: 273-278.

Puligilla, C., A. Dabdoub, S. D. Brenowitz and M. W. Kelley (2010). "Sox2 Induces Neuronal Formation in the Developing Mammalian Cochlea." The Journal of Neuroscience **30**(2): 714-722.
Read, G., D. Bedford and M. Feldman (1970). "Gastrula." Notes **27**(2): 339.

Rodaway, A. and R. Patient (2001). "Mesendoderm: An Ancient Germ Layer?" Cell **105**(2): 169-172.

Saga, Y. (1998). "Genetic rescue of segmentation defect in MesP2-deficient mice by MesP1 gene replacement." Mechanisms of Development **75**(1-2): 53-66.

Saga, Y., N. Hata, S. Kobayashi, T. Magnuson, M. F. Seldin and M. M. Taketo (1996). "MesP1: a novel basic helix-loop-helix protein expressed in the nascent mesodermal cells during mouse gastrulation." Development (Cambridge, England) **122**(9): 2769-2778.

Saga, Y., N. Hata, H. Koseki and M. M. Taketo (1997). "Mesp2: a novel mouse gene expressed in the presegmented mesoderm and essential for segmentation initiation." Genes & development **11**(14): 1827-1839.

Saga, Y., S. Kitajima and S. Miyagawa-Tomita (2000). "Mesp1 expression is the earliest sign of cardiovascular development." Trends in cardiovascular medicine **10**(8): 345-352.

Saga, Y., S. Miyagawa-Tomita, A. Takagi, S. Kitajima, J. i Miyazaki and T. Inoue (1999). "MesP1 is expressed in the heart precursor cells and required for the formation of a single heart tube." Development (Cambridge, England) **126**(15): 3437-3447.

Sakurai, H., T. Era, L. M. Jakt, M. Okada, S. Nakai, S. Nishikawa and S.-i. Nishikawa (2006). "In vitro modeling of paraxial and lateral mesoderm differentiation reveals early reversibility." Stem cells (Dayton, Ohio) **24**(3): 575-586.

Seo, S., J. W. Lim, D. Yellajoshyula, L. W. Chang and K. L. Kroll (2007). "Neurogenin and NeuroD direct transcriptional targets and their regulatory enhancers." The EMBO Journal **26**(24): 5093-5108.

Shen, X., B. Soibam, A. Benham, X. Xu, M. Chopra, X. Peng, W. Yu, W. Bao, R. Liang, A. Azares, P. Liu, P. H. Gunaratne, M. Mercola, A. J. Cooney, R. J. Schwartz and Y. Liu (2016). "miR-322/-503 cluster is expressed in the earliest cardiac progenitor cells and drives cardiomyocyte specification." Proceedings of the National Academy of Sciences **113**(34): 9551-9556.

Skinner, M. K., A. Rawls, J. Wilson-Rawls and E. H. Roalson (2010). "Basic helix-loop-helix transcription factor gene family phylogenetics and nomenclature." Differentiation **80**(1): 1-8.

Smale, S. T. (2010). "Pioneer factors in embryonic stem cells and differentiation." Current opinion in genetics & development **20**(5): 519-526.

Soibam, B., A. Benham, J. Kim, K. C. Weng, L. Yang, X. Xu, M. Robertson, A. Azares, A. J. Cooney, R. J. Schwartz and Y. Liu (2015). "Genome-Wide Identification of MESP1 Targets Demonstrates Primary Regulation Over Mesendoderm Gene Activity." STEM CELLS **33**(11): 3254-3265.

Soleimani, V. D., H. Yin, A. Jahani-Asl, H. Ming, C. Kockx, W. van Ijcken, F. Grosveld and M. A. Rudnicki (2012). "Snail Regulates MyoD Binding-Site Occupancy to Direct Enhancer Switching and Differentiation-Specific Transcription in Myogenesis." Molecular Cell **47**(3): 457-468.

Stern, C. D. (2005). "Neural induction: old problem, new findings, yet more questions." Development: 2007-2021.

Strahl, B. D. and D. C. Allis (2000). "The language of covalent histone modifications." Nature **403**(6765): 41-45.

Takahashi, Y., S. Hiraoka, S. Kitajima, T. Inoue, J. Kanno and Y. Saga (2005). "Differential contributions of Mesp1 and Mesp2 to the epithelialization and rostro-caudal patterning of somites." Development (Cambridge, England) **132**(4): 787-796.

Technau, U. and C. B. Scholz (2002). "Origin and evolution of endoderm and mesoderm." The International journal of developmental biology **47**(7-8): 531-539.

Teif, V. B. and K. Rippe (2009). "Predicting nucleosome positions on the DNA: combining intrinsic sequence preferences and remodeler activities." Nucleic Acids Research **37**(17): 5641-5655.

Thiery, J. and J. P. Sleeman (2006). "Complex networks orchestrate epithelial–mesenchymal transitions." Nature Reviews Molecular Cell Biology **7**(2): 131-142.

Trapnell, C., B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold and L. Pachter (2010). "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation." Nature Biotechnology **28**(5): 511-515.

Tuchmann-Duplessis, H., G. David, P. Haegel, H. Tuchmann-Duplessis, G. David and P. Haegel (1971). "Embryogenesis." springer: 34-41.

Tuchmann-Duplessis, H., G. David, P. Haegel, H. Tuchmann-Duplessis, G. David and P. Haegel (1971). "Embryogenesis." springer: 30-33.

van den Ameele, J., L. Tiberi, A. Bondue, C. Paulissen, A. Herpoel, M. Iacovino, M. Kyba, C. Blanpain and P. Vanderhaeghen (2012). "Eomesodermin induces Mesp1 expression and cardiac differentiation from embryonic stem cells in the absence of Activin." EMBO reports **13**(4): 355-362.

Venkatesh, S. and J. L. Workman (2015). "Histone exchange, chromatin structure and the regulation of transcription." Nature Reviews Molecular Cell Biology **16**(3): 178-189.

Wamstad, J. A., J. M. Alexander, R. M. Truty, A. Shrikumar, F. Li, K. E. Eilertson, H. Ding, J. N. Wylie, A. R. Pico, J. A. Capra, G. Erwin, S. J. Kattman, G. M. Keller, D. Srivastava, S. S. Levine, K. S. Pollard, A. K. Holloway, L. A. Boyer and B. G. Bruneau (2012). "Dynamic and Coordinated Epigenetic Regulation of Developmental Transitions in the Cardiac Lineage." Cell **151**(1): 206-220.

Wang, X., A. Spandidos, H. Wang and B. Seed (2012). "PrimerBank: a PCR primer database for quantitative gene expression analysis, 2012 update." Nucleic Acids Research **40**(D1).

Wegner, M. (2011). "SOX after SOX: SOXession regulates neurogenesis." Genes & Development **25**(23): 2423-2428.

Wells, J. M. and D. A. Melton (1999). "VERTEBRATE ENDODERM DEVELOPMENT." Annual Review of Cell and Developmental Biology **15**(1): 393-410.

Wild, A. E. and T. P. Fleming (2002). "eLS." wiley.

Wood, H. (2002). "Sox, flies and neurogenesis." Nature Reviews Neuroscience **3**(9): 675-675.
Wu, S. M. (2008). "Mesp1 at the Heart of Mesoderm Lineage Specification." Cell Stem Cell **3**(1): 1-2.

Yao, Z., A. P. Fong, Y. Cao, W. L. Ruzzo, R. C. Gentleman and S. J. Tapscott (2013). "Comparison of endogenous and overexpressed MyoD shows enhanced binding of physiologically bound sites." Skeletal Muscle **3**(1): 1-9.

Zhang, Y., T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li and S. X. Liu (2008). "Model-based Analysis of ChIP-Seq (MACS)." Genome Biology **9**(9): 1-9.

Zheng, X., X. Zheng, Y. Wang, Y. Wang, Q. Yao, Q. Yao, Z. Yang, Z. Yang, K. Chen and K. Chen (2009). "A genome-wide survey on basic helix-loop-helix transcription factors in rat and mouse." Mammalian Genome **20**(4): 236-246.

Zorn, A. M. and J. M. Wells (2009). "Vertebrate Endoderm Development and Organ Formation." Cell and Developmental Biology **25**(1): 221-251.